TECHNISCHE
UNIVERSITÄT
DARMSTADT

IMPROVING ONLINE PRIVACY AND SECURITY THROUGH CROWDSOURCED TRANSPARENCY PLATFORMS AND OPERATOR NOTIFICATIONS

Vom Fachbereich Informatik
der Technischen Universität Darmstadt
genehmigte

DISSERTATION

zur Erlangung des akademischen Grades
Doktor-Ingenieur (Dr.-Ing.)
von

MAX JAKOB MAASS

Erstreferent: Prof. Dr.-Ing. Matthias Hollick
Korreferent: Prof. Dr. Dominik Herrmann

Darmstadt 2021
Hochschulkennziffer D17

SEEMOO
SECURE MOBILE NETWORKING

Max Jakob Maaß, *Improving Online Privacy and Security Through Crowdsourced Transparency Platforms and Operator Notifications*, Dissertation, Technische Universität Darmstadt, 2021.

Fachgebiet Sichere Mobile Netze
Fachbereich Informatik
Technische Universität Darmstadt
Jahr der Veröffentlichung: 2021
Tag der mündlichen Prüfung: 2. Juli 2021
URN: urn:nbn:de:tuda-tuprints-191903

ABSTRACT

Modern life relies on the internet for everything from communicating and shopping to banking and seeking medical advice. However, this growth of internet-based services also leads to a higher risk of security and privacy issues. Finding and remediating these issues is an important challenge which cannot be addressed through purely technical means, as legal, economic, and psychological factors can also play a role in how these issues are created and resolved. This dissertation approaches this challenge from two sides: we discuss how to collect data and detect issues in the web and email ecosystems, and how the operators of affected systems can be convinced to address them.

Today, efforts to understand internet ecosystems frequently rely on automated large-scale scans. These can efficiently investigate large numbers of systems, but cannot access some ecosystems that require manual actions (e.g., signing up for a newsletter or account). To gather research data and gain access to new ecosystems, we propose and develop two public transparency platforms for use by internet users which collect information about security and privacy issues in the web and email ecosystems using a crowdsourcing approach. We consult with legal experts to ensure the adherence of our platforms to the relevant legislation. Over the 4 years of operation the platforms collected over 3 million scan results, which can serve as a basis for future research.

Our platforms also revealed a number of privacy, security and compliance issues, which should be addressed by the operators of the affected systems. Past research has shown that notifying operators about issues and convincing them to make changes is a challenging problem and frequently results in unsatisfactory remediation rates. We thus investigate the factors influencing the success of large-scale notification campaigns. For this purpose, we conduct three notification studies that evaluate different methods to incentivize system operators to address the issues, like inducing a competitive pressure (leveraging our existing public platform), highlighting the security threat an issue poses, or informing the operators that their systems are not compliant with relevant legislation. We also evaluate the choice of the message medium and the sender as factors in the success of a notification campaign. We collaborate with researchers from economics, law, and psychology to gain additional insights into the behavior of organizations and individual operators. Finally, we derive organizational and methodological recommendations for future notification campaigns based on our experience.

ZUSAMMENFASSUNG

In unserem täglichen Leben nutzen wir das Internet für viele Zwecke, von der Kommunikation über den Einkauf bis zum Banking und der Suche nach medizinischem Rat. Die zentrale Rolle die das Internet für uns spielt bedeutet auch, dass es ein hohes Potential für Sicherheits- und Privatheitsprobleme hat. Diese zu finden und zu beheben ist eine wichtige Aufgabe, die nicht rein technischen addressiert werden kann: auch juristische, wirtschaftliche und psychologische Faktoren spielen eine Rolle darin, wie diese Probleme entstehen und behoben werden. Die vorliegende Dissertation betrachtet dieses Themengebiet von zwei Seiten: wir betrachten die Datensammlung und das automatische Auffinden von Problemen für Webseiten und Email-Newsletter, und wir prüfen, wie die Betreiber*innen dieser Systeme dazu gebracht werden können, die gefundenen Probleme zu beheben.

Versuche, den aktuellen Zustand verschiedener Internet-Ökosysteme zu verstehen, basieren häufig auf groß angelegten automatischen Scans. Solche Scans ermöglichen es, schnell eine große Anzahl von Zielen zu untersuchen, scheitern aber an Systemen die eine vorherige Anmeldung benötigen, wie beispielsweise E-Mail Newsletters. Daher entwickeln und beschreiben wir in dieser Dissertation zwei öffentliche Transparenz-Systeme, die mittels Crowdsourcing Sicherheits- und Privatheitsprobleme in Webseiten und E-Mail-Newsletters identifizieren. Die Ergebnisse werden anschließend aufbereitet und auf einer Webseite öffentlich angezeigt. Dabei stellen wir in Zusammenarbeit mit juristischen Expert*innen sicher dass der Betrieb dieser Plattformen mit geltendem Recht verinbar ist. In den letzten vier Jahren haben Nutzer*innen über diese Plattformen über 3 Millionen Untersuchungen angestoßen und damit einen wertvollen Datensatz für zukünftige Forschungsvorhaben gesammelt.

Durch diese Plattformen ist es uns ebenfalls gelungen, einen Datensatz von Privatheits-, Sicherheits- und Konformitätsprobleme auf Webseiten zu sammeln. Bisherige Forschungsergebnisse haben gezeigt dass ein einfacher Hinweis an die Betreiber*innen häufig nicht ausreichend ist, um die Probleme beheben zu lassen. Daher untersuchen wir Faktoren die den Erfolg oder Misserfolg einer solchen Benachrichtigungskampagne beeinflussen können. Insgesamt führen wir drei Benachrichtigungsstudien durch, mit denen wir verschiedene Methoden untersuchen, Betreiber*innen dazu zu bewegen, die Probleme zu lösen. Zu diesen Faktoren gehören unter anderem der Wettbewerb mit anderen Webseiten, eine detaillierte Information über die Risiken eines Sicherheitsproblems, oder der Hinweis auf geltendes Recht, welches die Webseite aktuell verletzt. Desweiteren betrachten wir den Einfluss, den das Medium und der Absender einer Nachricht haben.

Diese Studien entstehen in Zusammenarbeit mit Forscher*innen aus der Wirtschaftsinformatik, Jura und Psychologie, um die Reaktionen aus der Perspektive der jeweiligen Disziplinen zu untersuchen. Wir schließen diese Dissertation mit einer Reihe organisatorischer und methodischer Empfehlungen für zukünftige Benachrichtigungsstudien, die wir aus unseren Erfahrungen ableiten.

ACKNOWLEDGMENTS

They say it takes a village to raise a child, and the same is true for a dissertation. I would not be where I am today if it wasn't for the support of many others. In particular, I wouldn't even have considered pursuing a PhD if it wasn't for Prof. Dominik Herrmann (who showed me that research is fun), Dr. Haya Shulman (who told me I was good enough), Prof. Marc Fischlin (who showed me the joy of teaching others) and of course Prof. Matthias Hollick (who encouraged me and my friends to write a paper based on a lab project and funded our trip to the conference). Thank you for setting me on the path for this adventure, whether you knew it or not.

I am further indebted to Prof. Hollick and Prof. Herrmann for (formally and informally) supervising my dissertation and supporting me through the many iterations and changes of my research ideas. I am grateful for the support of my colleagues, be it through proofreading, discussions, technical help, or encouragement. In no particular order: Mikhail Fomichev, Daniel Wegemer, Jiska Classen, Max Engelhardt, Lars Almon, Arash Asadi, Milan Stute, Matthias Gazzari and Annemarie Mattmann all helped me in their own ways. Outside of our research group, I thank Henning Pridöhl and Pascal Wichmann for their technical knowledge and collaboration, and Anne Laubach, Nora Wessels, Jacqueline Brendel, Kris Shrishak, Alina Stöver, Sebastian Bretthauer, Dirk Müllmann and Prof. Indra Spiecker for sharing their knowledge of cryptography, economics, law, and psychology. I am also grateful to the more than 50 students who I supervised in labs, seminars, bachelor-, and master theses. This experience has taught me a lot, and many of them left a mark on my research with their own contributions.

My friends helped me stay sane during my PhD. Lennart, Paula, Mareike, Sebi, Niels, Afra, Nora, Daniel, Max (the other one), Matthias, Jiska, Kris, Alina, Matthias (yes, another one), and Annemarie: thank you for our talks, board games, pen-and-paper rounds, dinners, walks, help with moving (twice!), movie nights, music recommendations, and everything else. The same is also true of my parents, who supported me every step of the way. I would also like to take this opportunity to thank the two cats that kept visiting my balcony during the isolation of the COVID-19 pandemic, loudly insisting that I take a break: Stella and Bailey, thank you.

Finally, I am grateful for the funding of the two research projects that paid my bills and let me meet all the interesting researchers who helped put this research together: the interdisciplinary Research Training Group 2050 "Privacy and Trust for Mobile Users" and the National Research Center for Applied Cybersecurity ATHENE. I am also deeply appreciative of the thousands of volunteers who develop and maintain the Open Source software that I use every day and without which all of my research would have been impossible: Thank you for all you are doing.

CONTENTS

LIST OF FIGURES

LIST OF TABLES

LISTINGS

ACRONYMS

CA	Certification Authority
CDN	Content Distribution Network
CERT	Computer Emergency Response Team
CMS	Content Management System
CSIRT	Computer Security Incidence Response Team
CSP	Content Security Policy
CSR	Corporate Social Responsibility
DDoS	Distributed Denial of Service
DKIM	DomainKeys Identified Mail
DNS	Domain Name System
EU	European Union
FIRST	Forum of Incident Response and Security Teams
GDPR	General Data Protection Regulation
HSTS	HTTP Strict Transport Security
HTML	Hypertext Markup Language

IA	Institutional Approach
IDS	Intrusion Detection System
ISP	Internet Service Provider
IXP	Internet eXchange Point
MX	Mail eXchange
NGO	Non-Governmental Organization
OS	Operating System
PCI DSS	Payment Card Industry Data Security Standard
RbV	Resource-based View
SPF	Sender Policy Framework
SSH	Secure Shell
TLD	Top-Level Domain
TLS	Transport Layer Security
VCS	Version-Control System
XSS	Cross-Site Scripting

LIST OF PUBLICATIONS

During the course of writing this thesis, I co-authored the following papers and articles.

JOURNAL AND MAGAZINE ARTICLES

[1] Mikhail Fomichev, Max Maass, Lars Almon, Alejandro Molina, and Matthias Hollick. "Perils of Zero-Interaction Security in the Internet of Things." In: *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 3.1 (2019). Received the Distinguished Paper Award of UbiComp 2020.

[2] Mikhail Fomichev, Max Maass, and Matthias Hollick. "Zero-Interaction Security - Towards Sound Experimental Validation." In: *GetMobile: Mobile Computing and Communications* 23.2 (2019).

[3] Milan Stute, Max Maass, Tom Schons, Marc-André Kaufhold, Christian Reuter, and Matthias Hollick. "Empirical insights for designing Information and Communication Technology for International Disaster Response." In: *International Journal of Disaster Risk Reduction* (2020).

CONFERENCE AND WORKSHOP PAPERS

[4] Santiago Aragon, Marco Tiloca, Max Maass, Matthias Hollick, and Shahid Raza. "ACE of Spades in the IoT Security Game: A Flexible IPsec Security Profile for Access Control." In: *2018 IEEE Conference on Communications and Network Security (CNS)*. IEEE. 2018.

[5] Steffen Klee, Alexandros Roussos, Max Maass, and Matthias Hollick. "NFCGate: Opening the Door for NFC Security Research with a Smartphone-Based Toolkit." In: *14th USENIX Workshop on Offensive Technologies (WOOT 20)*. 2020.

[6] Max Maass, Marc-Pascal Clement, and Matthias Hollick. "Snail Mail Beats Email Any Day: On Effective Operator Security Notifications in the Internet." In: *16th International Conference on Availability, Reliability and Security (ARES 2021)*. **Part of this thesis.** 2021.

[7] Max Maass, Anne Laubach, and Dominik Herrmann. "PrivacyScore: Analyse von Webseiten auf Sicherheits- und Privatheitsprobleme - Konzept und rechtliche Zulässigkeit." In: *INFORMATIK 2017 Workshop "Recht und Technik"*. **Part of this thesis.** 2017.

[8] Max Maass, Henning Pridöhl, Dominik Herrmann, and Matthias Hollick. "Best Practices for Notification Studies for Security and Privacy Issues on the Internet." In: *3rd International Workshop on Information Security Methodology and Replication Studies (IWSMR '21)*. **Based on this thesis.** 2021.

[9] Max Maass, Stephan Schwär, and Matthias Hollick. "Towards Transparency in Email Tracking." In: *Annual Privacy Forum*. **Part of this thesis.** 2019.

[10] Max Maass, Alina Stöver, Henning Pridöhl, Sebastian Brethauer, Dominik Herrmann, Matthias Hollick, and Indra Spiecker. "Effective Notification Campaigns on the Web: A Matter of Trust, Framing and Support." In: *30th USENIX Security Conference*. **Part of this thesis.** 2021.

[11] Max Maass, Nicolas Walter, Dominik Herrmann, and Matthias Hollick. "On the Difficulties of Incentivizing Online Privacy through Transparency: A Qualitative Survey of the German Health Insurance Market." In: *Wirtschaftsinformatik 2019*. **Part of this thesis.** 2018.

[12] Max Maass, Pascal Wichmann, Henning Pridöhl, and Dominik Herrmann. "PrivacyScore: Improving Privacy and Security via Crowd-Sourced Benchmarks of Websites." In: *Annual Privacy Forum*. **Part of this thesis.** 2017.

[13] Milan Stute, Max Maass, Tom Schons, and Matthias Hollick. "Reverse Engineering Human Mobility in Large-scale Natural Disasters." In: *Proceedings of the 20th ACM International Conference on Modelling, Analysis and Simulation of Wireless and Mobile Systems*. ACM. 2017.

POSTERS AND DEMONSTRATORS

[14] Max Maass, Uwe Müller, Tom Schons, Daniel Wegemer, and Matthias Schulz. "NFCGate: an NFC relay application for Android." In: *Proceedings of the 8th ACM Conference on Security & Privacy in Wireless and Mobile Networks*. ACM. 2015.

DATASETS

[15] Mikhail Fomichev, Max Maass, Lars Almon, Alejandro Molina, and Matthias Hollick. *Data and Code for "Perils of Zero- Interaction Security in the Internet of Things"*. Zenodo, 2019. URL: https://doi.org/10.5281/zenodo.2537721.

[16] Max Maass, Marc-Pascal Clement, and Matthias Hollick. *Data and Code for "Snail Mail Beats Email Any Day: On Effective Op-*

erator Security Notifications in the Internet". Zenodo, 2021. URL: https://doi.org/10.5281/zenodo.4817464.

[17] Max Maass, Alina Stöver, Henning Pridöhl, Sebastian Bretthauer, Dominik Herrmann, Matthias Hollick, and Indra Spiecker. *Data and Code for "Effective Notification Campaigns on the Web: A Matter of Trust, Framing, and Support"*. Zenodo, 2020. URL: https://doi.org/10.5281/zenodo.4075131.

SOFTWARE

[18] Max Maass, Alexandros Roussos, Daniel Wegemer, Steffen Klee, Tom Schons, and Uwe Mueller. *NFCGate Codebase*. URL: https://github.com/nfcgate/nfcgate.

[19] Max Maass, Stephan Schwär, Jonathan Schad, Jan Klinkmann, and Chi Viet Vu. *PrivacyMail Codebase*. URL: https://github.com/PrivacyMail/PrivacyMail.

[20] Henning Pridöhl, Pascal Wichmann, Dominik Herrmann, Max Maass, Martin Müller, and Malte. *PrivacyScore Codebase*. URL: https://github.com/PrivacyScore/PrivacyScore.

[21] Tom Schons, Milan Stute, and Max Maass. *Natural Disaster Mobility Models Codebase*. URL: https://github.com/seemoo-lab/natural-disaster-mobility.

UNDER PEER REVIEW

[22] Matthias Gazzari, Annemarie Mattmann, Max Maass, and Matthias Hollick. "My(o) Armband Leaks Passwords: An EMG and IMU Based Keylogging Side-Channel Attack." In: *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* (2021).

[23] Dirk Müllmann and Max Maass. "Online Evidence Collector – Ein Onlinetool zur einfachen und rechtssicheren Erstellung elektronischer Beweise." In: *Juristenzeitung (JZ)* (2021).

COLLABORATIONS AND MY CONTRIBUTION

In the following, I detail the contributions of my co-authors and myself per chapter. In addition, I follow the regulations of the Department of Computer Science at Technische Universität Darmstadt and give an account of the parts that include verbatim or revised fragments of previous publications that form this dissertation as indicated in the preceding list of publications.[1]

Chapter 1 and 2 summarize the contributions, background and related work of the core papers of this dissertation [6, 7, 9–12].

Chapter 3 is based on joint work with Dominik Herrmann, Henning Pridöhl, Pascal Wichmann and Anne Laubach [7, 12]. The underlying platform was written by Dominik Herrmann, Henning Pridöhl, Pascal Wichmann and myself, and has since been expanded and partially rewritten by Pascal Wichmann and Henning Pridöhl. My focus was on the data analysis aspects of the platform. The legal expertise in [7] was provided by Anne Laubach. Everyone contributed equally to writing the papers.

Chapter 4 is based on joint work with Stephan Schwär [9], who wrote significant parts of the backend as part of his Master thesis, and Jonathan Schad, Chi Viet Vu and Jan Klinkmann, who rewrote parts of front- and backend later. I participated in coming up with the system design, and wrote the frontend for the first prototype version, which has since been rewritten as part of Jonathan Schad's Bachelor thesis and a later lab project by Jonathan Schad, Chi Viet Vu and Jan Klinkmann. I wrote significant parts of the paper. Parts of the chapter, primarily between Section 4.1 and Section 4.3, are adapted verbatim or with small changes from the original paper.

Chapter 5 is based on joint work with Nicolas Walter, Dominik Herrmann, and Nora Wessels [11]. Nicolas Walter conducted the study and the first evaluation of the data as part of his Master thesis, co-supervised by Nora Wessels and myself. I subsequently re-evaluated the data with new techniques and wrote the paper jointly with Dominik Herrmann and Nicolas Walter. Parts of the chapter, primarily between Section 5.3.3 and Section 5.6 as well as all figures and tables, are adapted verbatim or with small changes from the original paper.

Chapter 6 is based on joint work with Marc-Pascal Clement [6], who conducted the study and the first version of the evaluation as part of his Bachelor thesis. Once again, I re-evaluated the data using new techniques and wrote significant parts of the paper, with input from Marc-Pascal Clement. The figures and tables are adapted verbatim or with small changes from the paper. The descriptions from Appendix

1 References in this chapter refer to my list of publications given on Pages xvii to xix.

B are partially adapted with changes from Marc-Pascal Clement's Bachelor thesis.

Chapter 7 is based on a collaboration with Alina Stöver, Henning Pridöhl, Sebastian Bretthauer, Dominik Herrmann, and Indra Spiecker [10], based on an idea primarily developed by Dominik Herrmann, Henning Pridöhl and myself. Henning Pridöhl developed the scanning infrastructure and self-service tool used in the study, and wrote the evaluation of the tool. Alina Stöver designed, conducted and evaluated the survey and worked with me to evaluate the received messages. Sebastian Bretthauer and Indra Spiecker provided legal expertise and fielded phone calls as part of the study. I prepared and sent the messages, responded to questions from recipients, and performed the survival analysis and literature work. We jointly wrote the paper, with each person writing the parts that fell inside their area of expertise. The figures and tables as well as parts of Appendix C were adapted verbatim or with small changes from the paper or its supplementary material.

Chapter 8 is based on experiences made in the previous three studies. It was written by me, with input from Henning Pridöhl and Dominik Herrmann. It was later used as a basis for a paper with the same authors [8].

Part I

INTRODUCTION

We begin this dissertation with an introduction to two internet ecosystems and the problems they face. We then give an overview of the pertinent literature from computer science and other relevant fields.

1 INTRODUCTION

In the last 30 years, the internet has seen an explosive growth in the number of connected devices, websites, services and business models. Services that were unthinkable 20 years ago are a routine part of the lives of millions of people today, and more and more of our lives is conducted online. However, this is not without risk.

1.1 MOTIVATION

Ensuring that the infrastructures powering our lives are secure, trustworthy, and respecting our privacy has become a critical task. Unfortunately, this goal has proven hard to achieve in practice. To understand the issues, we highlight three different perspectives on the ecosystems that make up the internet: that of the *end users* using its services, of the *operators* that build and maintain them, and of the *researchers* that want to understand the current state of the ecosystems.

Security and privacy are difficult to ensure.

END USERS End users are by far the weakest participants in the system, lacking any direct way to influence the security and trustworthiness of the services they are using, while having to trust them with their data, knowingly or unknowingly. The slew of successful attacks on major companies and vendors ranging from credit scoring[1] to social networks[2] have demonstrated that many companies are far from achieving the necessary level of security to warrant this trust. Instead, data breaches exposing the information of millions of users have become so commonplace that they hardly generate sustained attention, except in the most egregious cases.

The security and privacy of end users' data is threatened by insecure services.

At the same time, these companies threaten the privacy and security of their users through their use of third-party services. The most well-known of these are the advertising and profiling companies, which routinely collect and trade information about hundreds of millions of users, profile them for targeted advertising, and sell access to the resulting profiles, which are in turn exploited for (sometimes deceptive or manipulative) advertising [20].

Their privacy is also under threat from tracking companies.

Most of these issues are invisible to the users, who frequently do not know how many tracking and advertising companies are present on a

1 See https://arstechnica.com/?post_type=post&p=1166391, last accessed 2021-01-20.

2 See https://www.cbsnews.com/news/linkedin-2012-data-breach-hack-much-worse-than-we-thought-passwords-emails/, last accessed 2021-01-20.

website. This lack of transparency leads to system operators[3] suffering no ill effects (in terms of reputation loss or competitive disadvantages) from violating the privacy of their users, as these are unlikely to "vote with their feet" and take their business to more privacy-conscious competitors if they cannot easily compare them [57].

Tracking and security issues are invisible to the user.

It may be tempting to lay the blame for these practices at the feet of the system operators. However, we will see that they also face a number of constraints.

SYSTEM OPERATORS Modern websites are complex systems with many moving parts and external dependencies, ranging from web frameworks or Content Management Systems (CMSs) to libraries providing specific functions, cloud/hosting providers, and firewall systems. Any of these components may fail or contain vulnerabilities, introducing threats to the availability and security of the system. Even worse, these external dependencies may in turn have their own external dependencies, further increasing the exposure to these problems. Keeping track of this expanding web of dependencies is a significant challenge, leading to issues such as the use of outdated and vulnerable software, incompatibilities, and misconfigurations. The system may also end up violating relevant regulations like the General Data Protection Regulation (GDPR), exposing its operators to legal liability.

Modern web development is a complex process with many moving parts, which causes problems to system operators.

Many services' business models rely, at least in part, on advertising, either by showing ads themselves or by using advertising to attract customers. Under these circumstances, continuing to use tracking and advertising systems (in websites, apps, or even email newsletters) is often the economically correct decision, as the externalities are shouldered by the users. If this calculation is to be changed, the cost of using invasive tracking and advertising services must outweigh their benefit — and where this cannot be ensured by users or competitors, it may be achievable by regulation. But to understand in which areas regulation may be required, regulators need to understand the current practices and problems of internet ecosystems. Here, researchers may be able to provide valuable data.

Modern business models rely on advertising and tracking.

RESEARCHERS In the last 10 years, great steps have been made in improving the technical capabilities for large-scale scanning. Tools like ZMap[4] [33] allow scanning the entire IPv4 address space for a specific open port in less than an hour, and systems like Shodan[5] have large databases of internet services and their fingerprints. Researchers have also performed large-scale scans of websites [1, 2, 27, 38, 39, 42, 62,

Researchers have developed excellent tools for large-scale internet scanning.

3 For the purpose of this dissertation, we use the umbrella term "system operator" for both the *technical* (i.e., administrators) and *legal* operators (i.e., owners of the company).

4 See https://zmap.io, last accessed 2021-01-20.

5 See https://shodan.io, last accessed 2021-01-20.

114], and similar datasets exist for Android apps [52, 90, 95], among other areas.

Large-scale datasets face issues of validity and may neglect some ecosystems.

Given this wealth of data and technology, detecting vulnerabilities and misconfigurations at scale seems to be simple. However, such large datasets come with their own problems. For example, some question the validity of using common lists of popular websites (most commonly the Alexa Top Million) as datasets for such scans, citing issues in replicability and the large churn of these lists [64, 94]. It also leads to research favoring ecosystems that can be automatically analyzed with relative ease, which leads to less attention being paid to other ecosystems that are more challenging to investigate.

Notifying operators of insecure systems has proven challenging.

Collecting such large datasets also raises a challenging question: what should be done if the scans reveal systems that are vulnerable to outside attacks? This leaves researchers in a bind: while notifying all affected system operators results in a large workload for the researchers, *not* notifying them is ethically questionable. In practice, notifications often end up being undeliverable [17, 18, 34, 68, 98, 99], and the researchers are sometimes faced with hostile responses and even legal threats [17] if they send notifications, further complicating this decision.

1.2 CHALLENGES AND GOALS

Our research focuses on detecting problems and ensuring they are remediated by system operators.

In this dissertation, we seek to understand the current state of different internet ecosystems and determine how to influence operators to improve the privacy and security of their systems. In practice, these seemingly simple goals hide a high degree of complexity that can pose formidable challenges. Entire dissertations have been written about just a single facet of these issues [36, 67]. We thus begin by discussing the challenges and goals that we investigate in this dissertation, before highlighting our concrete contributions in the next section.

1.2.1 *Understanding Ecosystems*

Internet scanning suffers from issues with access and representativeness.

While large-scale internet scans can provide valuable insights, they also suffer from some limitations. In some ecosystems, data or software is not made available in an easily accessible way, as it may require payment or a manual sign-up (e.g., email newsletters). In these cases, manual work and/or spending money is required, which limits the scalability of the analysis. In some cases, it may also be difficult to find a representative sample of systems for the analysis.

PUBLIC TRANSPARENCY TOOLS We may be able to mitigate these issues by creating public transparency tools that are geared towards both system operators and end users. Such tools can be used to collect a different type of dataset compared to the one-time large-scale scans.

Public scanning tools can help with data collection and serve as a basis for future studies.

By collecting data about systems specifically requested by end users, such tools are no longer bound to the commonly used sources of target systems, and can thus avoid reproducing the same structural biases.[6] They can also be used to outsource manual work involved in data collection (e.g., signing up for newsletters) to the end user, which reduces the required time investment for the researcher. Finally, such platforms can build their own user bases, which may be helpful for follow-up studies (e.g., user studies or any study that relies on having a well-known platform in the background).

Building tools for under-researched ecosystems can be beneficial for both users and researchers.

NEW ECOSYSTEMS Much of the previous research has focused on the web [1, 2, 27, 38, 39, 42, 62, 114] or mobile app [52, 90, 95] ecosystems. Other ecosystems have received comparatively little attention, and some have little to no tool support for users or operators. Closing this gap can allow any tools developed by researchers to attract users and collect unique datasets that can help advance our understanding of these ecosystems.

1.2.2 *Changing Ecosystems*

To remediate detected issues, the cooperation of system operators is required.

While automated tools can reveal issues, remediating them requires system operators to act. This can be thought of as a two-step process: first, the system operators need to be successfully contacted so that awareness about the issue can be raised. Secondly, they need to actually remediate the issue by implementing changes to the systems under their control.

It is difficult to find contact information for system operators and gain their trust.

REACHING SYSTEM OPERATORS Although finding contact information for the operators of a website can usually be achieved within a minute of browsing the website, this approach does not scale to the thousands or tens of thousands of websites that are routinely found to be vulnerable on the internet. However, no repository of contact information for all websites exists, which leads to high rates of undeliverable messages [17, 18, 34, 68, 98, 99], especially after many Top-Level Domains (TLDs) removed address information from their automated WHOIS interfaces in response to the GDPR [70]. And even if a valid email address can be correctly guessed [96], recipients (rightly) distrust unsolicited messages [15, 16, 98, 117]. Thus, gaining enough trust to have the message taken seriously is another challenge that has to be overcome.

FACTORS INFLUENCING REMEDIATION Once a message has been received, read, and not dismissed out of hand, one might be tempted to assume that the problem will be addressed. However, a final chal-

6 Of course, any dataset created by such a tool carries its own biases based on the user base, which need to be considered in any evaluation of the data.

lenge remains: ensuring that the system operator will actually take action. Several studies have shown that awareness of a problem is not necessarily sufficient to ensure that remediation will be attempted, or completed successfully [17, 34, 68]. This can have several reasons, including a lack of understanding about the severity of the reported issue, lack of time to implement a fix, or even disagreeing with the premise that it actually constitutes a problem worth addressing [68]. Thus, it is important to determine which factors influence the willingness of system operators to attempt remediating an issue. At the same time, even if remediation is attempted, it may not be successful. System operators thus also need to be supported in validating the success of their remediation attempts.

Even if the operators can be reached, they may not remediate the issue, or remediate incompletely.

1.3 CONTRIBUTIONS

To address these challenges, we make three contributions: (1) we design automated public transparency tools for two internet ecosystems, (2) we evaluate factors that influence operators' willingness to make changes to their systems in response to outside notifications, while (3) sharing expertise and collaborating with researchers from other disciplines.

This thesis makes contributions in three areas.

1.3.1 *Automated Transparency Tools for Two Ecosystems*

To allow for the collection of different datasets and serve as a basis for further research, we develop two public transparency tools: PrivacyScore.org and PrivacyMail.info. These tools cover the web and email ecosystems, respectively. PrivacyScore considers privacy and security issues, and is being widely used (between 250 and 500 visitors per day as of May 2021, with over 2.9 million scans performed). The data it collects has also been used in subsequent studies [82, 93, 97], both ours and from other institutions. PrivacyMail.info scans email newsletters for tracking technologies. It is a younger platform and thus has fewer users (50-150 users per day, with over 350 000 emails analyzed) and follow-up research, but it fills a gap that no other public platform has addressed before: a long-term dataset of commercial newsletters which can be analyzed for privacy-invasive technologies. At the time of writing, the latter dataset has been shared with two other research projects, one of which resulted in a publication [54].

We develop public tools for the web and email ecosystems that are seeing widespread use.

In summary, we make the following contributions:

- We design and build PrivacyScore.org, an open source[7] transparency system that detects privacy and security issues on websites (Chapter 3).

7 See https://github.com/privacyscore/privacyscore, last accessed 2021-03-01.

- We design and build PrivacyMail.info, an open source[8] transparency system that evaluates email newsletters for trackers and other privacy issues (Chapter 4).
- We operate both as public services and share data from both platforms with other researchers to enable further research (Section 3.5 and 4.5).

1.3.2 *Changing Ecosystems*

We conduct three notification campaigns and synthesize lessons for future studies.

To investigate the factors that influence if operators will change their systems in response to outside notifications, we conduct three notification studies in which we evaluate different incentives and other factors that influence system operators' willingness to make changes to their websites. These factors include the contact medium, the message sender, and the content of the message. We also synthesize the procedural and technical lessons we have learned while conducting these studies into a set of best practices and recommendations for future studies.

In summary, we make the following contributions:

- We conduct a notification study to evaluate whether competition can serve as an incentive to improve online privacy in a notification study with 152 health insurance companies (Chapter 5).
- In a second notification study with 1359 website operators, we investigate the influence of providing more explicit descriptions of attacks enabled by detected security issues (Chapter 6).
- We perform a third notification study with 4594 website operators to investigate the effect of different senders and compare the remediation rates of a legal (compliance) argument with a baseline privacy argument (Chapter 7).
- We also use the latter two studies to compare the effectiveness of postal notifications with email notifications (Chapter 6 and 7).
- To facilitate future research, we collect the lessons we learned while conducting these studies and derive a set of best practices for future notification studies (Chapter 8).

1.3.3 *An Interdisciplinary Approach*

We collaborate with other disciplines to deepen our understanding.

Written as part of the DFG-funded interdisciplinary research training group "Privacy and Trust for Mobile Users", this dissertation benefits from collaborations and discussions with a large group of researchers

8 See https://github.com/privacymail/privacymail, last accessed 2021-03-01.

from diverse disciplines, ranging from law and economics to psychology and sociology. Discussions with the other PhD students as well as the postdocs and professors inform this work in many areas, and have resulted in several collaborations which allow us to provide a more holistic view of the topic areas discussed in this dissertation.

The following areas particularly benefit from these collaborations:

- We consult with researchers from the field of Information Systems to understand the perspectives of companies on privacy and security issues (Section 2.4) and evaluate competition as a factor in driving change to websites (Chapter 5).
- We work with law researchers to evaluate the legal questions surrounding the operation of our transparency systems (Section 3.4).
- We collaborate with a different legal research group to construct and evaluate a compliance argument and use their chair as one sender in our compliance notification study (Chapter 7).
- We work with researchers from psychology to analyze the responses of notification recipients in our compliance study (Chapter 7).

1.4 OUTLINE

We summarize the structure of the dissertation.

The rest of this dissertation is structured as follows: first, we provide required background information and an overview of related work in Chapter 2. In Part ii we describe the two automated transparency systems we developed, PrivacyScore.org (Chapter 3) and PrivacyMail.info (Chapter 4). We then utilize the developed technology for three notification studies in Part iii, discussing the use of competition (Chapter 5), alternative contact channels (Chapter 6), and alternative senders and non-technical arguments (Chapter 7) as potential factors that may influence remediation. We conclude the dissertation in Part iv by discussing the organizational and methodological lessons we learned from our notification studies in Chapter 8 before summarizing the high-level takeaways in Chapter 9.

2

BACKGROUND AND RELATED WORK

In this chapter, we give a background on different internet ecosystems and how they can be analyzed for issues. We then discuss the prior research on large-scale notification studies that sought to reach system operators and inform them about issues with their systems so that these can be remediated. Finally, we consider some basic theories from the field of economics that can help us to understand the view of organizations on security and privacy issues, and the role that transparency can play in changing their calculus.

This chapter gives an overview of relevant prior research.

2.1 INTERNET ECOSYSTEMS

The internet consists of a large combination of different ecosystems, including websites, apps, email, messengers, and many others. For the purpose of this dissertation, we consider two different ecosystems: Websites (or, shorter, "the web") and email. We thus briefly introduce these two ecosystems and the challenges they face.

We begin by considering two internet ecosystems: The web and email.

2.1.1 *The Web Ecosystem*

The web has become a central part of the lives of many people. It mediates our contact with friends (via social networks), financial transactions, entertainment, the search for information, and increasingly our work as well. It consists of a vast and ever-growing network of websites and services, run by countless small and large companies, individuals, and non-profit organizations. This decentralized nature allows the web to be a dynamic environment, quick to react to changes, but it also leads to a distribution of responsibilities and risks. We highlight two of these risks in more detail below.

The web is a highly dynamic environment, which leads to risks.

2.1.1.1 *Security*

Modern websites are frequently built with either pre-made CMSs or built on top of web frameworks like Express.js and Django, using popular libraries like Bootstrap, jQuery, and React. They also outsource complex functions like customer help chats, media playback, newsletter subscriptions, analytics and advertising to external companies. The website is then hosted, often with a large hosting company that operates powerful Content Distribution Networks (CDNs). This structure leads to several potential security issues.

Websites rely on common frameworks and a variety of services provided by external actors, which can prove a security risk.

Common frameworks lead to common vulnerabilities.

CENTRALIZED VULNERABILITIES Using a common CMS or framework allows rapid development of websites using an existing ecosystem of extensions and documentation and makes it likely that a website can be kept compatible with modern browsers and features with a comparatively small effort. However, this benefit comes at the cost of creating large numbers of websites running the same code and thus being affected by the same bugs and vulnerabilities. If a vulnerability is found that can be automatically exploited at scale, it can put the entire ecosystem at risk. This makes the timely installation of software updates critical.

External services are implicitly trusted, but may be compromised.

EXTERNAL DEPENDENCIES Using external services to provide functionality reduces development time and costs. However, it also leads to the website loading (and implicitly trusting) code developed by outside parties. If such a company is compromised, this can in turn compromise all websites that use their services.[1] Alternatively, the company may go out of business, which can lead to their domain becoming available again. If criminals subsequently register the same domain, they can begin distributing malware or other harmful content to all websites that still embed content from the domain.[2]

Operators may misconfigure their systems and unintentional expose sensitive information.

MISCONFIGURATIONS Even if the code running on a website is secure, the website may still be incorrectly configured to expose sensitive data like internal configuration details, Version-Control System (VCS) repositories[3], and even database backups [98]. Such misconfigurations can be found through automated scans and easily exploited at scale, and can stay hidden from the operator unless they perform vulnerability scans of their own website.

Finally, security may be compromised by vulnerabilities in the server software or intermediary systems.

INSECURE INFRASTRUCTURE Finally, even if the code is secure and the configuration is correct, the website may still be exposed to vulnerabilities in the web server, like the infamous Heartbleed attack [34]. These vulnerabilities may also lay outside of their direct control to remediate, as is the case with insecure Domain Name System (DNS) servers [17] or cache services[4]. Such issues may put the website and its users at significant risk, and are usually impossible to predict in advance.

1 See, for example, https://scotthelme.co.uk/hardening-payment-forms-with-csp/, last accessed 2021-02-02.

2 See https://0xpatrik.com/subdomain-takeover-basics/, last accessed 2021-01-14.

3 See https://en.internetwache.org/dont-publicly-expose-git-or-how-we-downloaded-your-websites-sourcecode-an-analysis-of-alexas-1m-28-07-2015/, last accessed 2021-04-01.

4 See https://blog.cloudflare.com/incident-report-on-memory-leak-caused-by-cloudflare-parser-bug/, last accessed 2021-02-02.

2.1.1.2 *Privacy*

Operators and users have different interests in the area of privacy.

Security issues like those mentioned above usually have the interests of users and operators aligned, as both would prefer the website to be secure. In the area of privacy, this is not necessarily the case. Here, the interest of the website operator may clash with the desire for privacy by the users.

Tracking is desireable for operators, but violates the privacy of the users.

TRACKING Many websites are operated with the goal of selling goods or services to its users. For these websites, it is desireable to determine how the visitors use the websites, as this allows website operators to determine which routes through the website the users take, and thus also gain insight into which aspects of the website drive purchases and which detract from the overall performance. Such analytics features are offered as a service by many companies, the most well-known of which is Google with its *Google Analytics* service. These systems collect significant amounts of data about the behavior of the users, in some cases collecting all of their actions down to mouse movements and keystrokes for later playback [1].

Advertising can keep services free for users or attract customers, but has significant privacy implications.

ADVERTISING AND PROFILING Closely related is the issue of advertising and profiling: some websites earn money not through the sale of products, but by using free content or services to attract users and then placing advertisements to monetize the attention. The online advertising ecosystem relies heavily on privacy-invasive profiling [20] to determine which ads to show to which user, and uses tracking to determine the effectiveness of an advertising campaign. This functionality is provided by dedicated third-party services that aggregate information about the behavior of the user from many websites to construct detailed profiles.

Operators must comply with privacy regulation, which can prove challenging.

REGULATION AND COMPLIANCE The extensive data collection of advertising and tracking companies has led to a pushback, most notably in the form of new regulations that limit the collection and use of personal information. In the European Union (EU), the most notable of these regulations is the GDPR.[5] Companies must comply with these regulations or face significant fines. However, many companies either struggle with compliance due to unclear and conflicting information about the legality of certain practices, or seek to circumvent the rules through psychological tricks like *dark patterns* that aim to obtain "informed consent" from the user by making the withholding of consent

5 While the EU ePrivacy directive, Digital Services Act, and Digital Markets Act promise to introduce further changes, none of them has been passed by the EU legislature at the time of writing.

difficult. This practice has been challenged in court[6], but no binding final ruling of the European courts has been made yet.

2.1.2 *The Email Ecosystem*

The email ecosystem has a different threat model than the web.

Where the web ecosystem provides a large variety of services, emails are more limited. They are mostly used for personal or professional communication, automated confirmations of orders, and newsletters. Both sender and recipient(s) presumably have an interest in keeping the message confidential. However, the question is: confidential against whom? Depending on the answer, different threats and technologies become relevant.

2.1.2.1 *Confidentiality against Network Providers*

Messages can be protected against untrusted networks through transport encryption, but it is frequently disabled.

The first and most simple case is that user and sender want to keep the content of their message confidential from the network it travels over, but have no interest in protecting against a curious mail provider. In these cases, emails can be protected in transit using a transport encryption technology like Transport Layer Security (TLS). However, in practice, even this simple form of protection is surprisingly rare [32, 55, 77], exposing many emails to snooping by anyone with access to the network traffic while they are being transmitted between mail servers.

2.1.2.2 *Confidentiality against Mail Providers*

Protecting against the email server requires end-to-end encryption, which is rarely used.

The next higher level of protection would be to protect the message against inspection by the operator of the sending and receiving email accounts. In this case, the encryption needs to be performed on the device of the sender, with the message being decrypted by the recipient (end-to-end encryption), using a technology like PGP or S/MIME. This requires special software for both parties, making its use rare in the general population.

2.1.2.3 *Confidentiality against Sender*

The message sender may also attempt to track the recipient.

Finally and perhaps counterintuitively, the recipient may want to protect themselves against the sender. The sender may want to know if and when the recipient opened the email and if they clicked any of the embedded links. As emails are often sent in the Hypertext Markup Language (HTML) format, they can load remote content (like images hosted on external servers). Such an image load can be detected by the server. Thus, embedding a personalized image link into each email can

6 See https://www.hiddemann.de/allgemein/lg-rostock-bejaht-unterlassungsanspruch-bei-nudging-ueber-cookie-banner/ for a German description of such a case. Last accessed 2021-02-02.

allow the sender to know when the email was read, and personalized links can be used to detect when a link was clicked. Both techniques are commonly used in commercial mailings [37, 46, 50, 54, 113].

Email tracking can disclose sensitive information that cannot easily be obtained by other forms of tracking.

Email tracking is arguably even more powerful than website-based tracking: The recipients are already linked to a persistent identifier that will change only very rarely — their email address — which allows tracking their behavior over long timeframes. It also permits advertisers to easily link the behavior from multiple newsletters in a single profile, further increasing the value of the data. If users click a tracked link from within an email, the newsletter-based profile can be linked with a web-based profile identified by browser cookies. Email is also frequently used from more than one device (i.e., laptop, phone and tablet). Thus, if links are clicked from more than one device, the profiles collected from the two devices can also be linked to the same person [37].

No good defenses against email tracking exist.

There are also significantly fewer options for users to defend themselves from tracking — if they are even aware that email tracking exists, which they frequently aren't [113]. Dedicated email clients only rarely offer ad-blockers, and even if they are installed, the block lists are frequently insufficient [37, 54, 113]. And even though tracking through remote content can be avoided by not loading this content, this frequently renders the email newsletter unreadable due to missing design elements. Finally, tracked links are also impossible to avoid, unless the user wants to manually search for the linked article or product on the website, further reducing the convenience of the newsletter. Thus, email tracking is a powerful and hard-to-defeat threat to the privacy of users.

2.2 AUTOMATED ECOSYSTEM ANALYSIS

We discuss two approaches for ecosystem analysis.

In order to determine how wide-spread specific privacy and security issues are, automated analyses can prove helpful. Ecosystems can be analyzed using two methods: automated large-scale scans can give a view of the overall state of the environment, while public transparency tools can provide more targeted information about individual systems (while potentially still collecting a high-level picture from the aggregated scans). We discuss both types in more detail below before considering their relative strengths and weaknesses.

2.2.1 *Large-Scale Automated Scanning*

Large-scale scans provide an overview of an ecosystem.

Large-scale scanning systems rely on easily automated and scalable processes. Their intent is not to investigate or promote changes to a specific system, but rather to gain a high-level overview of the current state of the ecosystem, looking for aggregate results and trends without special consideration for individual systems. This information

can then be used to determine how prevalent specific issues are, inform policy decisions, or serve as a baseline for more detailed investigations.

Such scans are supported by many existing tools and services.

In this section, we briefly consider the use of large-scale scans in two ecosystems which are of interest for this dissertation — the web and email. However, large-scale automated scanning is also used in many other areas, with services like Shodan[7] or software like ZMap[8] [33] providing the infrastructure for detailed scans of non-HTTP internet services.

2.2.1.1 *The Web Ecosystem*

Large-scale scans are commonly used in the web.

Large-scale Web scans have become so accessible that they can be conducted with relative ease, even by small teams or individuals. Thus, both the security and privacy aspects of the web have been analyzed with large-scale scans.

Security scans have revealed many issues on the web, and are supported by existing services and datasets.

SECURITY In the area of security, scans are used to detect issues as diverse as outdated software [62], information leaks[9], or (mis)use of TLS [49]. Services like Censys[10] [31] or Crawler.ninja[11] collect regular scans of large swaths of the internet, with results detailing the use of TLS, Security Headers or certification information. This allows even those without the ability to perform such scans to work with the results and perform analyses.

Large-scale privacy scans are more complex, but are supported by existing software as well.

PRIVACY In the area of privacy, scans are frequently complicated by the fact that many privacy-relevant issues can only be detected by using a fully functional browser (instead of simple scripted HTTP requests) to load and execute the JavaScript of the website and capture all connections it establishes. This leads to significantly increased resource requirements to perform such scans at scale. Tools like OpenWPM [38] package a full browser with instrumentation code that allows remotely controlling its actions. Such tools can then be used to conduct studies about web tracking [1, 38], the use of cookies [39], or the prevalence of *dark patterns* [76].

2.2.1.2 *The Email Ecosystem*

The email ecosystem is less suited for large-scale scans.

In the email ecosystem, large-scale scans are less common — some aspects can be evaluated easily, while others require the semi-automated collection of a large corpus of emails, which can prove challenging. We

7 See https://shodan.io, last accessed 2021-01-20.
8 See https://zmap.io, last accessed 2021-01-20.
9 See https://en.internetwache.org/dont-publicly-expose-git-or-how-we-downloaded-your-websites-sourcecode-an-analysis-of-alexas-1m-28-07-2015/, last accessed 2021-04-01.
10 See https://censys.io/, last accessed 2021-01-20.
11 See https://crawler.ninja/, last accessed 2020-01-26.

thus discuss large-scale scans divided by the three different protection goals discussed above.

CONFIDENTIALITY AGAINST NETWORK PROVIDERS The availability of transport encryption can be evaluated using automated scans that probe common email-related ports. Thus, several studies exist that investigate this question [32, 55, 77] and usually come to sobering results: even though it is reasonably easy to deploy, many email servers still do not support transport encryption.

Scans for transport encryption are relatively straightforward.

CONFIDENTIALITY AGAINST MAIL PROVIDERS Due to the decentralized nature of email encryption systems, there is no feasible way of quantifying its use with large-scale scans. Some studies used the public PGP keyservers to perform basic analyses of the web of trust [107] or types of users [11] (using the keyservers as a directory of users that they contacted for a survey). The latter study shows that the users skew towards a male (94.9 % of respondents) and young population (almost 50 % below 35 years old), living in Europe or North America (combined 90 %) and working in IT or education (combined 78.1 %) [11]. This shows that usage of PGP (as measured by this study) is limited to a very specific demographic.

End-to-end encryption cannot be quantified by large-scale scans.

CONFIDENTIALITY AGAINST THE SENDER Researching the presence of tracking in emails requires having access to a large corpus of messages, which can be obtained from different sources. Previous studies used data collected from personal email accounts [113], disposable email services [50], or from commercial newsletters to which they signed up using an automated crawler [37]. This diversity of methodologies also highlights a major challenge of the field: unlike websites, email newsletters cannot be easily crawled, and their collection cannot be arbitrarily sped up through parallelization, as they depend on the frequency with which a publisher is sending their newsletters. Thus, large-scale studies ideally require large, diverse datasets that have been collected over long timeframes — a goal that may be easier to achieve through crowdsourcing using a public tool.

Scanning emails for tracking requires access to a large corpus of messages, which can be difficult to collect.

2.2.2 *Public Transparency Platforms*

Instead of untargeted mass scans, public transparency / scanning platforms rely on the interest of end-users to guide their scanning. Aside from their uses in large-scale ecosystem analysis through the aggregation of scan results, these platforms also provide a useful service to their users, which may be system operators or interested end users investigating the security or privacy of their favorite websites.

Providing public tools instead of large-scale scans can provide a different view on an ecosystem while also being useful to its users.

2.2.2.1 *The Web Ecosystem*

We list a small sample of tools for the web.

Many scanning platforms for websites exist. Instead of giving an exhaustive overview, which would be out of scope for this dissertation, we briefly highlight a few different systems from the areas of security and privacy as illustrative examples.

Security scanners are a common service on the internet, but are frequently geared towards operators instead of users.

SECURITY A large number of different scanners offer different types of security audits for websites. They range from specialized scanners for specific issues or systems (TLS security[12], HTTP security headers[13], etc.) to meta-scanners that consider several issues at once, like the Mozilla Observatory[14] or Immuniweb[15] scanners. While some of these scanners are run as non-profit services, many are operated by for-profit companies that use them to advertise their own products or services. This also means that most of the scanners target system operators instead of end users as their audience, and thus lack explanations of the results that would be understandable to a non-technical audience. This increases their usefulness to system operators, but significantly reduces it for end users.

Privacy scanners are less common, with many users preferring browser addons that also provide protection.

PRIVACY Privacy-focused scanners are much less wide-spread, and usually operated by non-profits (as is the case for the *Webbkoll*[16] platform [4]) or journalists (in the case of *Blacklight*[17], which is run by The Markup). More popular than these services are browser addons like PrivacyBadger[18], Ghostery[19], Disconnect[20], TrackingObserver[21] [92, Sect. 2.6] and Lightbeam[22] that can detect third-party tracking and block and/or visualize it for the users. Finally, some companies offer automated scans that check for compliance with the GDPR. However, these are usually paid services that are geared towards commercial customers and not end users.

2.2.2.2 *The Email Ecosystem*

The email ecosystem is less well supported with tools.

In the email ecosystem, only comparatively few tools exist, which are almost exclusively geared towards system administrators. Most are

12 See https://www.ssllabs.com/ssltest, last accessed 2021-01-26.
13 See https://securityheaders.com, last accessed 2021-01-26.
14 See https://observatory.mozilla.org, last accessed 2021-01-26.
15 See https://www.immuniweb.com/websec, last accessed 2021-01-26.
16 See https://webbkoll.dataskydd.net, last accessed 2021-01-26.
17 See https://themarkup.org/blacklight, last accessed 2021-01-26.
18 See https://privacybadger.org, last accessed 2021-01-26.
19 See https://www.ghostery.com, last accessed 2021-01-26.
20 See https://addons.mozilla.org/en-US/firefox/addon/disconnect/, last accessed 2021-01-26.
21 See https://trackingobserver.cs.washington.edu/, last accessed 2021-01-26.
22 See https://addons.mozilla.org/en-US/firefox/addon/lightbeam-3-0/, last accessed 2021-01-26.

focused on protecting mail traffic against the network providers, and almost none focus on the mail provider or sender.

Scanners for transport encryption exist, but suffer from technical limitations and are geared towards administrators.

CONFIDENTIALITY AGAINST NETWORK PROVIDERS The scanners from SSL-Tools.net[23] or Immuniweb[24] can perform automated scans of arbitrary mail servers. However, they are geared towards system administrators and do not offer explanations for end users. Additionally, they rely on establishing many connections to the mail server to test different TLS configurations, which can lead to mail servers slowing down or even blocking their connections (*tarpitting*[25]).

A few tools also seek to provide information that is intelligible for users.

One exception is the My Email Communications Security Assessment (MECSA) platform[26], operated by the Joint Research Center of the EU. MECSA is a security assessment tool that also seeks to provide a basic interpretations of its results that is understandable to end users without a technical background. At the same time, it has also been used to collect crowd-sourced research data [55]. Finally, two extensions for the email client *Thunderbird* seek to visualize the transport encryption status[27] and the jurisdictions an email crossed in its path to the recipients[28], respectively.

End-to-end encryption has very little direct scanning tool support.

CONFIDENTIALITY AGAINST MAIL PROVIDERS As email encryption relies on a decentralized ecosystem of encryption keys (for both PGP and S/MIME), no tools exist that can reliably determine if a specific recipient is using it. The closest to such a tool are keyservers[29] where keys can be published, or standards like AutoCrypt[30] that embed information about the availability of encryption keys in emails.

Email tracking detection tools currently operate exclusively as browser addons for the GMail webmail interface.

CONFIDENTIALITY AGAINST SENDER To the best of our knowledge, only two tools exist that can detect email tracking and warn users about it: UglyEmail[31] and PixelBlock[32]. Both are built as browser add-ons, and only work for the Google Mail webmail system. This leaves significant portions of email users (i.e., everyone who is not using GMail or prefers to use a dedicated email client instead of webmail) with no tools to detect and block tracking in emails.

23 See https://ssl-tools.net/mailservers, last accessed 2021-02-01.
24 See https://www.immuniweb.com/ssl/, last accessed 2021-02-01.
25 See https://www.techopedia.com/definition/1722/tarpitting, last accessed 2021-02-01.
26 See https://mecsa.jrc.ec.europa.eu/, last accessed 2021-02-01.
27 See https://addons.thunderbird.net/en-US/thunderbird/addon/paranoia/, last accessed 2021-02-01.
28 See https://tracemail.eu, last accessed 2021-02-01.
29 See http://pool.sks-keyservers.net/, last accessed 2021-02-01.
30 See https://autocrypt.org, last accessed 2021-02-01.
31 See https://uglyemail.com, last accessed 2021-02-01.
32 See https://chrome.google.com/webstore/detail/pixelblock/jmpmfcjnflbcoidlgapblgpgbilinlem?hl=en, last accessed 2021-02-01.

2.2.3 *Comparison*

We compare the two described scanning paradigms.

After discussing the different paradigms for ecosystem analysis, we now discuss their relative advantages and drawbacks to highlight at which tasks they excel and for which the alternative would be better suited.

2.2.3.1 *Large-Scale Automated Scans*

Large-scale scanning is fast, cheap, and does not rely on having a large user base.

Large-scale scans have three core advantages compared to user-facing scanning platforms: They are fast, with their runtime (aided by almost unlimited potential for parallelization) measured in hours or days instead of months or years. They are cheap(er) to develop and operate, as they can focus on making their scans efficient and do not have to consider other issues like building a user interface. Finally, they do not require a large user base to generate their datasets.

However, it can only investigate public systems, may create biased datasets, only gives a snapshot view of an ecosystem, and cannot notify operators about issues.

These advantages come at a cost. First, large-scale scans can only investigate systems that are publicly accessible and encounter challenges when applied to areas like email, where the content is not public. Second, as these scans usually only consider a fixed list of systems provided by their operators, they often consider a limited class of systems: those that are popular enough to show up in internet toplists like the Alexa Top Million. These lists have some structural issues [64, 94] and it is unclear if results obtained on them are representative for the *long tail* of less popular websites and systems.[33] Third, a single scan only gives a snapshot view of the question, with no information about how the results change over time. Fourth, Vekaria *et al.* have shown that the home page of a website may contain different trackers than sub-pages of the same website [109], which can lead to incomplete information being collected when scanning only the home page of a website. Finally, while a large-scale scan can generate a large list of systems that have (sometimes critical) issues, they do not offer an integrated way to inform the operators of these systems, leading to complex and costly large-scale notification campaigns (cf. Section 2.3).

2.2.3.2 *Public Transparency Platforms*

Public tools can rely on users to perform manual actions, may collect a more diverse dataset, and can offer instructions on remediating detected issues.

Public transparency / scanning platforms can solve many of these issues. As the targets of their scans are specified by users, more manual work can be performed (like asking users to send an email in the case of the MECSA platform [55]), which can increase the quality of the data (or make the collection possible in the first place). This crowdsourcing can also lead to a more diverse dataset that also contains samples of the long tail of less-popular websites. If the platform is used by the

33 In fact, a previous study by Englehardt and Narayanan showed that even within the top-million sites, the results differ significantly depending on the popularity of the website [38].

operator of a system, it can also directly offer instructions on how an issue can be remediated, thereby removing the need for complex notification campaigns later down the line.

They also have research applications.

From a research perspective, operating a platform also offers possibilities for further research, including user studies about how specific issues are perceived or how scan results can be best communicated. The technical basis and infrastructure of the platform can also often be repurposed for large-scale scans if more data should be collected on a specific issue.

However, they rely on having an active user base and may have different biases in the data.

However, in return for these advantages, scanning platforms suffer from a number of drawbacks. As they rely on input from their users, the data collection is frequently much slower, and they run the risk that the platform never reaches a critical mass of users that would suffice for data collection. The crowdsourcing approach also leads to a dataset that, while it does not suffer from the biases of internet toplists, likely contains other and harder to quantify biases based on who is using the platform. This may include a bias towards specific types of websites, or geographic biases in the user population.

They also encounter technical challenges in the evaulation and operation.

The data collected by such a platform also poses some challenges in the evaluation. If the platform is developed over time, older results may use a different methodology than more recent ones, making it impossible to compare them directly. It also means that the results are collected over a longer stretch of time, which means that results for some systems may no longer be accurate. These factors have to be considered during the evaluation, further increasing its complexity. Finally, building and maintaining a platform is more costly than a one-off large-scale scan: in addition to the scanning infrastructure, the platform requires a front-end for users and needs to be maintained for longer timeframes, increasing maintenance and hosting costs. The system also needs to be more resilient to incorrect (or malicious) use. All these factors combine to make operating a public platform less appealing to researchers.

2.3 EFFECTIVE LARGE-SCALE NOTIFICATIONS

We summarize the current research into operator notifications.

Once scans have detected an issue with a system, the operators need to be notified so that they can remediate it. While this is easily done for a small number of systems, it quickly becomes a complex endeavor if the list of vulnerable systems grows.

Previous research has considered many aspects of notifications, which we discuss below.

The question of how system operators can be effectively and efficiently notified about issues has been studied in several notification studies over the last 10 years. These studies have considered areas as diverse as the security of websites [14, 18, 34, 69, 98, 99, 108, 117] or DNS servers [17], the remediation of DDoS amplifiers [16, 60, 68], and even end-user malware infections [15]. They varied factors such as the contact channel, sender of the message, and the information given in

		Direct			Indirect					Format				Other Factors					
Paper	Year	WHOIS	Alias	Manual	Monitoring	Hosting	CSIRT	Browser	ISP	Verbosity	Appearance	Translation	Framing	Restrictions	Status	Sender	Reminders	Tool	Survey
Vasek *et al.* [108]	2012	●				●				●									
Canali *et al.* [14]	2013		●																
Kührer *et al.* [60]	2014						●												
Durumeric *et al.* [34]	2014	●																	●
Çetin *et al.* [18]	2016	●				●				●						●			
Li *et al.* [68]	2016	●					●			●		●					●		●
Li *et al.* [69]	2016	●			●			●				●		●				●	
Stock *et al.* [99]	2016	●	●			●	●										●		
He *et al.* [47]	2016			◐											●				
Çetin *et al.* [17]	2017	●	●			●												●	●
Stock *et al.* [98]	2018	●	●	●							●					●			●
Çetin *et al.* [15]	2018								●					●					
Zeng *et al.* [117]	2019	●			●							●	●			●			●
Çetin *et al.* [16]	2019								●					●					
Zhuang *et al.* [118]	2020			◐											●				
Tang *et al.* [105]	2020			◐											●				

Table 1: Overview of notification aspects discussed by prior work. ◐: not explicitly stated but inferred from the paper.

the notification itself. They also evaluated the effectiveness of sending reminder messages and providing tools for operators to validate their remediation. Many studies also sought input from the contacted system operators through surveys [17, 34, 68, 98, 117], with varying degrees of success. In the following, we summarize the current state of research. We also give an overview of relevant papers in Table 1.

2.3.1 *Contact Channel*

Operators of affected systems can be notified via several different channels.

Choosing which contact channel to use is one of the most consequential decisions a notification campaign makes. This includes the notification medium, the contacted party, and the method for finding their contact information. The approaches used in prior studies can be roughly divided into two categories: direct contacts with the affected party and the use of intermediaries. We now consider these two categories in turn.

2.3.1.1 *Direct Contact Channels*

Direct contact channels seek to reach the operator directly, usually using email messages.

Direct contact is the most straightforward way to notify a system operator about an issue. The main challenge lies in identifying the correct email address to contact. Prior studies used two methods to automatically determine this address: domain- or IP-WHOIS data [17, 18, 34, 68, 69, 98, 99, 108, 117], which has to be provided by system operators when registering a domain or receiving an IP range assignment, and addresses generated based on RFC 2142 [14, 17, 98, 99], a standard for purpose-specific email aliases [24]. One study also tested the use of manually-collected address information [98].

Address information from WHOIS records is frequently incorrect and often unavailable since the GDPR came into effect.

WHOIS While a WHOIS record is available for every domain and IP address, the data is frequently incorrect or resolves to unmonitored email addresses. This can be either an unintentional misconfiguration, or an intentional spam-protection measure by the operator. They thus suffer from bounces[34] (in some cases for more than 50 % of sent emails [17, 18, 99]) and low email open rates. Additionally, after the GDPR came into effect, contact information was removed from the public WHOIS data for many TLDs and is only released upon request for civil claims or criminal proceedings[35], making its use as a contact channel for security notifications impossible.

34 The term *bounce* is commonly used to describe a message being rejected by the receiving mail server, which will usually respond with a message containing the reason for the bounce (e.g., the email address does not exist, the mailbox is full, etc.).

35 See https://www.denic.de/en/whats-new/press-releases/article/extensive-innovations-planned-for-denic-whois-domain-query-proactive-approach-for-data-economy-and/, last accessed 2020-11-09.

STANDARD ALIASES Given these problems with WHOIS data, researchers have evaluated the use of standard aliases as an alternative. RFC 2142 [24] defines a series of email aliases and what purposes they should be used for. This includes aliases such as `security@domain.tld`, `abuse@domain.tld` or `webmaster@domain.tld`, but also the general-purpose alias `info@domain.tld`. If a website operator observes this RFC, they should have created these aliases and routed them to the correct person inside their organization. In practice, studies have reported that the bounce rates are still significant, even if multiple aliases are notified [98, 99].

RFC 2142 defines a number of purpose-specific email addresses for domains.

To investigate the reachability of these aliases in more detail, Soussi *et al.* performed automated scans and found that, depending on the sample, only between 4 and 22.6 % of websites appeared to offer at least one relevant alias, and the most common alias was `abuse@domain.tld` [96].[36] Thus, notifications will reach only a minority of websites when using this approach, and even fewer are likely to take action.

Experiments have shown that these aliases are frequently not used by operators.

MANUAL ADDRESS COLLECTION The unsatisfactory performance of many automated notification schemes raises the question if the problem lies with the used communication channel, or if it is a deeper issue with notifications. To begin to answer this question, Stock *et al.* performed a small-scale (N = 364) notification campaign using manually-collected address information and a set of different communication channels [98]. In addition to emails, they used social media, phone calls, contact forms, and letters. They found that manual contact channels led to a slightly increased remediation rate compared to automated emails, but that the increase did not justify the significantly increased work and financial costs. We note that their results should not be directly compared with other studies, as participants were selected based on the fact that they did not react to a prior automated notification. Thus, in addition to the small sample size, it may suffer from self-selection and priming effects. Nevertheless, the results indicate that recipients who did not respond to an automated message may still be reached through manual channels as a fallback mechanism.

Manual address collection is rarely used. Prior research found that the benefits do not outweigh the costs, but was based on a small sample.

2.3.1.2 *Indirect Contact Channels*

Given the issues with direct contact attempts, researchers began to consider other options. They settled on four strategies that relied on intermediaries (i.e., third parties that may have better options to reach

36 These numbers were obtained without sending emails and thus do not consider if the email would actually have been read. The numbers cited above do not include mailservers that are configured to accept emails for arbitrary addresses. If these are added, the availability increases to between 11.6 and 41.6 %, which can serve as an upper bound.

the system operator): utilizing existing website monitoring services like the Google Search Console[37] [69, 117], contacting the hosting or network providers [17, 18, 99, 108], or cooperating with Computer Security Incidence Response Teams (CSIRTs)[38] [60, 68, 99]. One study also used more drastic measures, displaying browser warnings when users attempted to visit an infected website [69].

Indirect contact channels use intermediaries to deliver the notifications.

Additionally, the affected systems may be operated by end users. This could be the case if a misconfigured consumer router exposes internal devices of the network to the internet, which can be infected with malware or exploited for Distributed Denial of Service (DDoS) amplification attacks. In these cases, the Internet Service Provider (ISP) of the end user may be the only party that can establish contact. Some ISPs have established processes to notify their customers of such misconfigurations, which have been used in two studies [15, 16].

If the affected systems are operated by end users, indirect channels are usually the only avenue for contacting them.

MONITORING SERVICES Some system operators may choose to register for external monitoring services to be informed of problems with their websites. These services have the downside that not everyone is registered for them, but they offer a verified communication channel to the system operators that use them. Additionally, by signing up for such a service, system operators signify a desire to be informed about issues.

Monitoring services like the Google Search Console can provide a contact channel to system operators that have signed up for them.

Two studies leveraged one such service, the Google Search Console, for their notifications about compromised [69] or misconfigured [117] websites. Li *et al.* reported that messages sent via the Search Console, combined with Google search result annotations or browser warnings, led to improvements in remediation rates compared to unmessaged websites [69]. However, they do not explicitly compare Search Console messages with WHOIS messages. Zeng *et al.* report that their Search Console messages resulted in a small (statistically insignificant) improvement in remediation rates compared to the WHOIS messages [117].[39] Thus, somewhat surprisingly, there is no evidence that having a pre-existing contact channel improves remediation rates.

Prior research found no improvement compared to direct contact when contacting operators via the Google Search Console.

HOSTING PROVIDERS Many system operators do not actually operate the physical hardware that their systems are running on. Instead, they rent servers or specialized services from providers, which can host anything from DNS records to entire websites. These service providers usually offer a mechanism for handling complaints from

The companies hosting the affected servers may also serve as an intermediary to to reach the operators responsible for the system.

37 See https://search.google.com/search-console/about, last accessed 2021-01-20.

38 I use the term CSIRT instead of the older and more commonly known term "Computer Emergency Response Team (CERT)", as the latter is trademarked by the Carnegie Mellon University. The referenced papers use the older term.

39 To avoid biasing their results, they base their comparison on recipients of WHOIS messages that are also signed up for the Google Search Console. It is thus possible that the difference compared to the full WHOIS group would have been significant, however, this statistic is not reported in the paper.

third parties, which they either act on themselves or forward to their customers.

Prior research has shown that hosting providers frequently fail to forward notifications to their customers.

Several studies contacted providers to inform about malware infections [18, 108], insecure websites [99] or misconfigured DNS servers [17]. One of these studies notified about issues that had to be remediated directly by the provider [17], thus making the provider the correct point of contact instead of an indirect channel. In this case, they found that contacting the provider directly was more effective than contacting their customers or upstream network provider. Two others did not explicitly compare the effectiveness of notifying the provider compared to direct notifications to the customers [18, 108], making it impossible to make statements about the relative effectiveness of the channel. Stock *et al.* found that providers frequently did not forward the notification to their customers, which rendered such notification prone to failure [99]. Thus, contacting providers does not appear to be a good solution for many situations.

CSIRTs coordinate vulnerability disclosure and thus appear to be good intermediaries for notifications.

CSIRTs CSIRTs coordinate the response to computer security incidents. They exist on many levels, from companies to national and international coordinating associations like the Forum of Incident Response and Security Teams (FIRST). Many accept reports about vulnerabilities and coordinate the mitigation and disclosure process with reporter and vendor. Three prior studies worked with these institutions, addressing issues such as DDoS amplification [60, 68], misconfigured firewalls [68], or vulnerable web applications [99].

CSIRTs often do not forward security notifications, frequently rendering them no more (or even less) effective than direct contact channels.

Kührer *et al.* used a large outreach campaign that spanned several CSIRTs to raise awareness for misconfigured servers that could be used for DDoS amplification, reducing the population of vulnerable servers by over 90 % for some misconfiguration classes [60]. However, they did not attempt to contact system operators directly, so no comparison exists. Li *et al.* found that direct contact via WHOIS performed better than attempting to reach system operators via CSIRTs, as some of the latter did not forward the vulnerability reports to the system operators, stopping the notification process in its tracks [68]. The same problem was observed by Stock *et al.*, who found that while CSIRTs *did* lead to improved outcomes compared to direct channels, this was mostly due to a single, highly-successful organization [99]. Without this organization, the remediation rates would have been similar to those of the direct contact channels.

Operators may also be reachable by displaying browser warnings when accessing their websites.

BROWSER WARNINGS If a direct notification to the system operator is unsuccessful, it may be possible to reach them by displaying browser warnings when accessing or searching for their website. Even if this notification does not immediately reach the operator, they may be informed by their users, or notice a sudden drop in website traffic and investigate the source.

Li *et al.* evaluated this approach in a collaboration with Google (in the same study that also used the Google Search Console), and found that browser warnings were more effective than simply annotating the website in the Google search results [69]. The performance was further improved if the website operators were also explicitly contacted via the Search Console. This indicates that more invasive methods are better in "getting the attention" of recipients and lead to a quick remediation to avoid further disruptions of their website.

This disruptive notification channel performed well in one prior study.

QUARANTINE NETWORKS When consumers are operating a misconfigured, vulnerable or compromised device, finding a way to notify them is challenging. Here, the ISPs can play an important role, as they have contact information for their customers, and can also access and modify their internet traffic. Two studies worked with an ISP to evaluate the effectiveness of their notification scheme that relied on both direct email contact with the affected customers, and on placing them in a *quarantine network*, which denied the customers access to large parts of the internet unless they remediated the issue (or manually released themselves from quarantine) [15, 16].

End users may be reachable through their ISPs, which can deny them access to the internet until they remediate an issue.

Both studies found quarantine networks to be a potent tool in driving remediation. In a 2018 study, Çetin *et al.* found quarantine networks to be effective, leading to remediation rates of 69 % for first-time quarantine events [15]. They later conducted a second study in which they compared quarantines with simple email notifications (without quarantine) and found the latter to have a worse performance (75 % remediation rate, compared to 87 % for quarantined users and 53 % for the control group) [16]. This indicates that, like in the case of browser warnings, more disruptive notification methods are also more effective in driving remediation. However, they also lead to more complaints: in both studies, about 10 % of quarantined users called their ISP to complain, and 3 % threatened to cancel their contract. Thus, these disruptive measures also come at a higher cost for the sender.

These so-called quarantine networks performed well in driving remediation, but led to customer complaints.

2.3.2 *Senders*

Intuitively, another factor that may influence the trust afforded to a notification message may be the sender of that message, where name recognition may lead to a higher initial trust. This idea was evaluated by three prior studies. Çetin *et al.* sent notifications using three different senders: a private security researcher, a university group, and *StopBadware*, a well-known anti-malware organization [18]. In their previously discussed study, Zeng *et al.* sent part of their messages via the Google Search Console (with Google branding), while the rest of the messages were sent by an email account associated with UC Berkeley [117]. Stock *et al.* took a different approach and

The sender of the message may have an impact due to name recognition.

compared messages that appeared to come from a human with those that claimed to come from an automated system [98].

However, three separate studies found that the sender did not have an impact in practice.

Surprisingly, all three studies reported the influence of the message sender on remediation as small and, where this was reported, statistically insignificant — a counter-intuitive result that warrants further investigation, as no convincing explanation has been found.

2.3.3 *Format*

The wording and formatting of the notifications may also influence remediation rates.

Aside from the message sender and medium, researchers can of course also vary the contents of the notification message. Here, prior studies have investigated four factors: the verbosity of the message, appearance, the use of different framings of the problem, and translations of the message into the native language of the recipients.

Messages with more details perform better than short messages.

VERBOSITY Three studies investigated if more detailed messages improve remediation success. All three studies found that messages with more detailed led to higher remediation rates than shorter messages [18, 68, 108], even if the shorter messages contained a link to a website with further information [68]. Vasek *et al.* actually found that their shortest messages were statistically indistinguishable from not notifying at all [108]. This indicates that notification messages should contain enough actionable information that the recipients can understand and validate the problem without consulting external resources.

The visual design of the message did not make a difference in one study.

APPEARANCE Only one study varied the visual design of the notification message for the same sender: Stock *et al.* compared HTML emails (with a well-designed message following the corporate design of their university) with plaintext messages and found no significant differences in remediation [98]. They also found HTML messages to suffer from lower deliverability, with the reading rate dropping by 2-3 percentage points (from ~12 to ~9 %) compared to plaintext messages, which may be explainable by spam filters.

Translating the message into the language of the recipient either did not influence remediation rates or reduced them.

TRANSLATION All previous notification studies used English as the primary language for the notification. However, many websites are operated by individuals or companies outside of english-speaking countries. Thus, three studies translated notifications into the language of the recipient. The first study by Li *et al.* did not evaluate the effect of translating messages (all messages were translated) [69]. However, in their second study they found that translated messages actually observed *reduced* remediation rates compared to their English counterparts [68]. They attribute this to recipients not expecting to receive a translated message from a US university and suspecting it of being phishing or spam. The final study, by Zeng *et al.*, saw no differences in

remediation rates, positive or negative [117]. As their translated messages were sent by Google, they suspect that the recipients were less surprised about receiving a localized version of the message, which may have prevented a more negative impact.

FRAMING Intuitively, another factor that may influence the notification success is the framing of the problem. Zeng *et al.* investigated two different framings for their notifications about TLS misconfigurations [117]: in the *technical focus*, they described the technical implications of the misconfiguration (i.e., the browser being unable to verify the authenticity of the connection), while the *user focus* instead focused on the experience of the website users, naming issues like the increased threat of data tampering or harms to the reputation of the website. They observed no differences in remediation rate between the two framings. This may also be due to the fact that both messages contained the information that website visitors may see browser warnings that impede access to the website — a potent incentive for remediation that may overshadow the rest of the arguments in the message. We thus discuss the role of incentives next.

Different framings of the problem did not make a difference in one study.

2.3.4 *Incentives*

Intuitively, the easiest method to motivate a system operator to remediate an issue is to ensure it is in their own best interest to do so. While a certain amount of intrinsic motivation can be assumed (especially in cases where operators are notified about security issues), some studies augmented this with additional incentives.

Intrinsic motivation may be insufficient to ensure remediation and could be enhanced through further incentives.

ACCESS RESTRICTIONS We have already considered two forms of incentives while discussing the indirect communication channels. They took the form of either making it harder for customers of a website to actually reach it while it was insecure (e.g., through browser interstitials or warnings in the Google search results [69]), or, in the case of end user with infected devices, denying them access to large parts of the internet until they remediated the issue [15, 16]. In cases where such an incentive was not universally applied, but compared against a group without such an incentive, they increased remediation rates [16, 69].

Restricting access to the website or the internet at large was an effective incentive.

LEGAL LIABILITY A different form of incentive is the risk of legal liability for inaction. Here, Çetin *et al.* reported that phishing and malware distribution sites were often remediated more quickly if they targeted banking credentials than in other cases [18], leading them to propose that the risk of punitive legal action may be a powerful motivator as well. The role of legal requirements in driving remediation has otherwise remained unexplored, although Diop *et al.* have done

Anecdotal evidence suggests that legal liability may serve as an incentive.

some preliminary work to investigate potentially relevant laws and regulations [29].

STATUS COMPETITION Finally, we briefly highlight four studies that were not primarily designed as notification experiments, but provide an interesting view on possible incentives. The studies relied on competition between companies [47, 104, 105, 118] to incentivize remediation, creating rankings where companies were compared based on the outgoing volume of automated spam messages. They investigated this approach in different countries and industry sectors, and three of them explicitly notified the companies about the ranking [47, 105, 118]. The studies found that rankings (as a form of reputational sanctions) were effective at reducing the volume of outgoing spam compared to an unranked control group, indicating that they may be a potential mechanism in other areas as well.

Status competition proved effective in four prior studies.

2.3.5 *Reminders*

If a recipient does not react to a notification, it does not necessarily mean that they are not going to remediate. They may simply have missed (or dismissed) the first message, and may be receptive to a reminder message. Two prior studies investigated the effects of sending reminder messages to system operators that had not remediated a few weeks after the initial notification, and found conflicting results: while Stock *et al.* reported a small effect from sending a reminder [99], Li *et al.* observed no significant increase in remediation from the reminder message [68]. The question of the effectiveness of reminders has thus not been settled yet.

The role of reminder messages is unclear, with two studies finding conflicting results.

2.3.6 *Tool Support*

Since many of the issues the notifications are concerned with can be detected automatically, it is also possible to provide the recipients with an automated tool that they can use to see if their remediation attempt was successful. This is especially important as previous research has shown that recipients sometimes attempt remediation, but fail to adequately address the issue and remain affected, sometimes unknowingly [17, 34, 68]. Thus, providing a tool for validation may increase overall remediation rates.

It may be possible to improve remediation by providing operators with a tool to verify their remediation.

Previous studies also proposed that such tools may prove helpful [68] or reported recipients requesting tool support [117], but only two studies actually provided a tool that allowed notification recipients to verify if their remediation attempts were successful [17, 69]. Of the two, only Çetin *et al.* actually evaluated the effect of the tool on remediation by providing it to some recipients while withholding it from others, as part of a notification study involving misconfigured DNS servers.

Only one study evaluated this hypothesis and found that a tool did not increase remediation rates.

They found that while many recipients expressed a desire for a tool when asked, and found it helpful when it was offered, providing or withholding the tool did not have a significant impact on remediation rates [17].

Of course, a tool may also have a positive impact outside of the direct remediation rates: similar to the effect of sending more detailed notification messages [68], it may decrease the amount of support requests and increase the confidence of system operators. On the other hand, it may also reduce trust as it requires providing a link to an external website in the notification, which has been reported as a trust-inhibiting factor [98]. It may even be legally problematic if it allows scanning arbitrary websites for security issues, necessitating access control and introducing further complications in the study design. Researchers should consider whether providing a tool is worthwhile on a case-by-case basis.

Providing a tool may also have other effects, positive and negative.

2.3.7 *Surveys*

Several prior studies included surveys of the recipients [17, 34, 68, 98, 117], asking after their views about the notifications. Many of these surveys suffered from low response rates, with many reaching less than 100 respondents [17, 34, 68, 117]. In addition, the surveys may suffer from self-selection effects, as respondents will, on average, likely have a higher trust towards the notification messages (those that distrust the notification are less to answer surveys). Nevertheless, the responses can provide valuable insight on topics such as problem awareness, the acceptability of unsolicited notifications, and the perceived trustworthiness of the notifications.

Previous studies surveyed recipients about their views on notifications, frequently suffering from low response rates.

PROBLEM AWARENESS Surprisingly, many system operators were aware of the notified issue. Durumeric *et al.*, notifying about the *Heartbleed* issue, reported that all 17 respondents had heard about the vulnerability, and the remaining vulnerable systems were either not managed by the respondent, or had been missed in their own scans [34]. Çetin *et al.* similarly reported that 40 % of their 25 respondents had taken action before receiving the notification, but been unsuccessful [17]. 88 % of respondents reported intending to remediate in response to the notification.

Many respondents had already been aware of the issue they were notified about.

Li *et al.* reported that 46 % of their 57 respondents had been aware of the notified issue (firewall misconfigurations, DDoS amplification, and exposed industrial control systems) before receiving the notification, and 16 % had attempted to remediate already [68]. These high numbers may also be explained by the fact that some system operators disagreed with the assessments of the researchers that the detected issue was actually a misconfiguration, and characterized them as intentional. This likely contributed to the fact that only 63 % of them

In some cases, perceived misconfigurations were intentional.

reported intending to take action in response to the notification. Such disagreements with the researchers' assessments were also found by Zeng *et al.*, who found system operators citing compatibility concerns as reasons not to deactivate outdated and insecure TLS cipher suites [117]. These counterintuitive results indicate that awareness of a problem is not the only barrier to remediation.

Notifications were generally deemed desireable.

ACCEPTABILITY System operators generally found unsolicited notifications acceptable [17, 34, 68, 98], with approval rates exceeding 90 % in some cases [34, 98]. Çetin reported a low number of hostile responses, including one threat of legal action and a "rather unimaginative insult" [17, p. 11]. However, in all reported cases, positive and thankful responses outnumbered negative and hostile ones.

Many factors influence the perceived trustworthiness of a notification.

TRUSTWORTHINESS The main challenge inhibiting notification trustworthiness seems to be the high prevalence of spam and scam messages received by system operators [68, 98]. Respondents reported the use of external links and unexpected and unknown message senders as trust-reducing factors [98]. Interestingly, messages that were sent by a US university but translated to the local language of the recipients were perceived as less trustworthy than the untranslated messages [68]. In their study comparing WHOIS messages with those sent via the Google Search Console, Zeng *et al.* found that messages sent via the Search Console were rated as more trustworthy [117], with respondents noting that they remembered opting in to these messages, making them less unexpected than the unsolicited WHOIS messages sent to other recipients. However, as discussed in Section 2.3.2, this did not translate into large increases in remediation rate. Overall, more research is needed to understand which factors influence the perceived trustworthiness of notification messages.

2.4 ORGANIZATIONAL DECISION-MAKING

The remediation behavior of organizations is influenced by non-technical factors.

Notification studies seek to determine how messages must be designed and delivered to ensure that they are trusted by the recipient. However, even if a message is trusted, the recipient may still choose not to act upon it, as they may not consider the issue worth addressing based on their own calculus of costs and benefits. Additionally, many systems are operated not by private individuals, but by organizations. To understand what factors may influence an organizations' willingness to make changes to promote privacy and security, we need to consider how organizations make such decisions. Understanding these views and priorities can help us to design notifications that take them into account, thereby increasing the likelihood that an issue is addressed. We will thus give a brief overview of existing works on security and privacy from the perspective of organizations.

2.4.1 *Economics of Security and Privacy*

The field of economics offers theories about the security and privacy behavior of companies.

According to common economic theories, the views and strategies a company holds about security or privacy may differ depending on many factors, including their industry sector, competition, marketing or financial goals. Due to this range of possibilities, we consider security and privacy separately.

2.4.1.1 *Security*

Organizational behavior differs depending on who is bearing the cost of (in)action.

Companies can invest in security in several different ways. For the purpose of this dissertation, we limit ourselves to two: Securing their own systems against attacks, and ensuring that their systems cannot be used to attack others. Each of these options has a different economic calculus attached to it.

Organizations consider the cost of remediation compared to the likelihood and cost of negative consequences for inaction.

PROTECTING THEMSELVES Intuitively, investing money in protecting yourself against loss is a rational action. However, in an economic consideration, the necessary investment needs to be proportional to the potential risk, both in terms of the probability of the event occurring, and in terms of the damage the event would cause. For this purpose, they use models to trade off the costs and benefits of investing in security [10, 12].

Quantifying that likelihood and cost is not trivial due to missing information. Thus, notifications containing such information may be more effective.

The main challenge faced by companies is that the risk (and thus the potential return on investment to prevent it) can be difficult to quantify. Some metrics have been proposed to measure this information [10], but calculating them frequently requires access to proprietary data about the frequency of specific attacks or other risks [51]. This calculus implies that notifications may be more effective if they explicitly quantify the potential risks of the issue they are notifying about to aid the recipient in making a correct cost-benefit-analysis.

Insecure systems that harm others (but not the operator) are a negative externality.

PROTECTING OTHERS While it is hard to convince companies to invest money to protect themselves, it is even harder to convince them to invest it to protect others. An example of such a case is when the networks of companies are abused to send spam or phishing messages to others. The sending company has very low costs associated with sending the messages, while the recipients face significantly higher costs (some authors estimate that every dollar earned by spammers costs 100 dollars in lost productivity on the recipient side [89]). Sending spam is thus a *negative externality* [89], like carbon emissions, where the emitting party does not bear the full cost of the action, but transfers it to other companies, or society at large.

This leads to a situation that is commonly referred to as the *tragedy of the commons*, where a public good (like email, or our planet) is made unusable because no one has an incentive to invest in prevention if no one else does the same. Research into this area, under the umbrella

The resulting tragedy of the commons situation may be addressable through reputation systems and sanctions.

of *commons theory*, has shown that when modeling the situation as a multi-round game, collaboration can be established using a reputation mechanism, where parties that have not collaborated in the past can be sanctioned while collaboration is rewarded [81]. It has also been shown that participants are willing to face losses themselves to punish non-collaboration (*altruistic punishment*) [40].

For some parts of the internet, such reputation systems and sanctions have already been established with some success.

In the context of network providers, these sanctions could take the form of de-peering networks that refuse to take action against spam, isolating them and raising their costs [7]. However, establishing a (formal or informal) reputation system with (formal or informal) sanctions may not be practical in all situations, especially if no direct business relationships between the relevant companies exist. One of the most successful examples of such a reputation system is *certificate transparency* [63], a public directory of issued TLS certificates which seeks to catch misbehaving Certification Authoritys (CAs) in the TLS ecosystem. Beginning as a voluntary system, it has helped to detect many mis-issued certificates[40] and has since become mandatory for all CAs trusted by the Chrome browser[41]. Transparency systems were also used with some success by four academic studies investigating spam emitted by corporate networks [47, 104, 105, 118].

2.4.1.2 *Privacy*

There are only few economical theories on the privacy behavior of companies.

When it comes to privacy, the views and needs of companies have received surprisingly little attention, with only a few articles considering the question [8, 43–45]. Generally, companies face a conflict in the area of privacy: On the one hand, their customers desire privacy and don't want to disclose data unless necessary, while on the other hand, the company needs data to improve their products, generate revenue, measure customer retention, and stay competitive. At the same time, companies operate within a legal framework that forces them to conform to minimum standards or face fines.

We consider a subset of these theories and refer to extant literature for more details.

For the purpose of this dissertation, we consider the two theoretical frameworks offered by Greenaway and Chan [44]: the Institutional Approach (IA) and the Resource-based View (RbV). We also briefly consider the strategies companies use to balance data use and privacy, as described by Gerlach *et al.* [43]. For further reading, we refer the interested reader to a later paper by Greenaway *et al.* [45] discussing four different *privacy orientations*. Additionally, the papers by Gerlach *et al.* [43] and Bélanger and Crossler [8] both contain surveys of research into organizational strategies in the area of privacy.

40 See https://security.googleblog.com/2015/10/sustaining-digital-certificate-security.html or https://scotthelme.co.uk/extended-validation-not-so-extended/ for two examples, last accessed 2021-02-15.

41 See https://archive.cabforum.org/pipermail/public/2016-October/008638.html, last accessed 2021-02-15.

THE INSTITUTIONAL APPROACH (IA) Under the IA, the behavior of companies is seen as a "search for legitimacy"[42] [44, p. 176] under external pressures and social norms. Greenaway and Chan [44] distinguish two approaches: the *acquiescent approach* consists of compliance with the law and imitation of their peers, but does not seek to exceed the minimum requirements, while the *proactive approach* consciously seeks to exceed minimum requirements and uses clear communication to achieve leadership. However, proactive firms still constrain themselves and do not deviate too far from industry standards, to avoid attacking and undermining the basic claim to legitimacy of their industry.

The Institutional Approach sees organizations as actors seeking to gain and maintain legitimacy.

This results in two basic strategies for companies: basic compliance or proactively exceeding the required minimum.

If we assume companies to operate in this model, *acquiescent* companies can be best reached by framing the notified privacy issue as an issue of non-compliance with legal or industry standards. A *proactive* company could also be reached by highlighting the potential for further differentiation by improving its practices (e.g., by using a more privacy-friendly tracking provider or enabling additional privacy features).

Acquiescent and proactive companies can be influenced with different strategies.

THE RESOURCE-BASED VIEW (RBV) The RbV instead considers privacy as a resource that is managed by the company to achieve a competitive advantage [44]. Under this view, companies either pursue an *information focus* or a *customer focus*. Under an information focus, superior data analysis allows the company to be more innovative and offer a better product through better knowledge of their customers, at the price of reduced privacy. A customer focus instead seeks to gain the trust and loyalty of the customer. Crucially, this does not necessarily mean that less data is collected, but that the purposes are made clearer and the customer is given more control over the process. As a simplified example, Google and Amazon could be considered to pursue the information focus, while Apple's public commitment to privacy and user control follows the customer focus.

The Resource-based View sees companies as managing their resources for competitive advantage.

They either maximize information collection to produce a superior product, or maximize the trust of their users at the cost of some information fidelity.

A company operating with an information focus may be persuaded to change its practices if it is presented with alternatives that allow it to maintain its information collection while still improving the privacy of the users (e.g., by using a self-hosted tracking solution instead of a commercial tracking and profiling provider like Google Analytics). Companies with a customer focus may similarly be open to alternative solutions for their own data collection, and may even accept slightly reduced fidelity to protect the privacy of its users (like the use of differential privacy by Apple[43]).

Companies following either of these archetypes can be influenced through different strategies.

42 Greenaway and Chan distinguish different types of legitimacy, which have different meanings. Companies may pursue some, but not other forms of legitimacy. For details, see [44, p. 177].

43 See https://www.apple.com/privacy/docs/Differential_Privacy_Overview.pdf, last accessed 2021-02-15.

Companies can also use balancing strategies and decide on privacy issues on a case-by-case basis. In these cases, extant literature has identified a number of strategies.

BALANCING STRATEGIES The previous theories treat companies as a monolithic institution following an archetype in order to be able to make broader claims about corporate strategies. When it comes to the day-to-day decisions inside a company, a different class of strategies must be considered. Gerlach *et al.* [43] conducted interviews with practitioners and distilled three sets of strategies that are used to balance privacy concerns and corporate interests: surrogate tactics (finding alternatives to data collection), segmentation (collecting data only from some users), and transparency tactics (controlling if and how information about the data collection is disclosed and presented). Here, efforts can be made to influence which strategies are chosen through a variety of methods that highlight the benefits of one solution over another, be they technical, legal or reputational.

2.4.2 *Transparency*

Reputation systems and competition both rely on transparency.

Both reputation-based sanctions (in the case of security) and privacy-based competition require peers and customers to be aware of the behavior of other companies. Establishing (semi-)public knowledge about and comparability of the practices of different companies falls under the umbrella of *transparency*. Coming from the economic and management perspective, Parris *et al.* define transparency as "the extent to which a stakeholder perceives an organization provides learning opportunities about itself" [85, p. 233]. Stated less formally, it is the degree to which an organization is open to inspection from outside stakeholders like customers or the general public.

Transparency in its many forms is a widely used building block in several fields, including computer science.

Transparency can take many forms, including voluntary data releases by the organization itself, mandated disclosures, and external audits, to name a few. It has been discussed and used as a building block in many areas, including Corporate Social Responsibility (CSR) [65, 78], human-rights policy [53], and nuclear disarmament / non-proliferation [115]. In computer science, it can take the form of scans of websites [2, 38, 39] or penetration tests [21, 58, 101, 106], but also systems like the previously-discussed *certificate transparency* [63].

Given that transparency is seen as desirable, companies may be tempted to co-opt it.

As transparency is perceived as positive by many stakeholders, there is a risk that organizations will attempt to simulate transparency while practicing *in*transparency. In this dissertation, I will thus first discuss how the transparency ideal can be subverted before using these limitations to illustrate the requirements for an effective transparency regime.

2.4.2.1 *Limitations of Transparency*

Extant literature has criticized the shortcomings of transparency in practice.

Several authors have considered the downsides and limitations of transparency in practice. We highlight some concerns below, and refer the interested reader to the articles by Ananny and Crawford [3] (discussing accountability), Stohl *et al.* [100] (discussing visibility and

effective use), and Christensen *et al.* [19] (discussing transparency as a communicative practice) for a more detailed criticism of the concept and practice of transparency.

Organizations may strategically over- or under-disclose to distract from undesireable behavior.

TRANSPARENCY AS A SMOKESCREEN Organizations may selectively disclose information in one area to distract from another. For example, nutrition labels provide one type of information about a product, but a good performance on such a label may mask the undisclosed environmental and social impact of the production chain [19].

Too much transparency can be as effective as too little in hiding information.

Alternatively, organizations may disclose so much information that the recipient is unable to understand and process all of it, thereby hiding incriminating information [3]. This is called *strategic opacity* by Stohl *et al.* [100], who also describe the *transparency paradox*, in which higher amounts of information lead to less knowledge as it becomes impossible to comprehend.

Transparency places significant responsibility on the receiver of the information.

TRANSPARENCY AS A BURDEN The capabilities of stakeholders, like customers, to analyze and understand the data is thus an important limitation to transparency. Additionally, proponents of transparency also assume that stakeholders are actually *interested* in the information, will not misunderstand it, and are willing to change their behavior in response, none of which is a given [19, 30, 112]. For example, no amount of transparency about working conditions in the supply chain will change the behavior of a consumer who is not interested in them.

Such criticisms of transparency can also be found in the data protection discourse.

Ananny and Crawford also offer a more fundamental critique of this point: they argue that the ideal of transparency places too much responsibility on the consumers, thereby reproducing a neoliberal ideal of personal responsibility as a replacement for government oversight and intervention [3]. This concern mirrors critiques of the "informed consent" model of data protection with its unreadable [80] and frequently incorrect [84] privacy policies, which are their own form of transparency.

Transparency can only be effective if stakeholders have some way of exerting influence on the decisions of disclosing organization.

TRANSPARENCY WITHOUT LEVERAGE Implicit in the idea of transparency is that, if the stakeholders change their behavior in response to disclosed information, this will have an effect on the disclosing organization and serve as a motivation to change their own behavior in turn. This assumes that it is vulnerable to public shaming [3] or that the stakeholder has another form of leverage [53]. It also presupposes that a viable alternative *exists*, which raises the issue of (quasi)monopolies like Facebook that can operate on a "take it or leave it" basis [112].

2.4.2.2 *Effective Transparency*

We discuss the requirements for effective transparency.

These criticisms of transparency also in turn reveal how a transparency regime must be designed to be effective. For this, research has iden-

tified two core concepts: embeddedness and leverage, the latter of which also leads to the related concept of competition.

The transparency process and its consequences need to be embedded into the decision processes of stakeholders and the organization itself.

EMBEDDEDNESS Embeddedness refers to information users (the person or entity making use of the disclosed information) and information subjects[44] *embedding* the information produced by the transparency process into their decision-making [112]. For example, customers may decide to stop purchasing from a company after specific information about its practices have come to light. This lack of new purchases is noticed by the company (i.e., the data is embedded in its decision making process), causing it to change its behavior.

The preparation of information for easier consumption can be outsourced to dedicated intermediaries.

In practice, the work of analyzing the information published by the discloser may be outsourced to an intermediary (or "infomediary" [30]), which eases the embedding into the decision process of the user by disseminating, controlling, verifying and translating the information for easier consumption [30]. Such infomediaries may take the form of Non-Governmental Organizations (NGOs) or consultancies that specialize in developing and auditing CSR policies, and selling their auditing reports to investors that implement ethical screening procedures for their investments [30]. Alternatively, a state-mandated transparency system can be designed to make complex information salient and accessible to the stakeholders, like publicly posted restaurant hygiene grading cards that contain a clear overall rating [112].

Public scanning services may be able to serve as such an intermediary.

From a computer science perspective, a successful scanning platform (cf. Section 2.2.3.2) can be seen as a kind of infomediary. While establishing such a platform is comparatively easy, ensuring that it becomes popular enough to actually have the effect of an infomediary is significantly more difficult, although some companies are attempting to establish themselves in such a role.[45]

Transparency must be combined with a plausible method for stakeholders to exert influence.

LEVERAGE If stakeholders have no plausible way of exerting influence over the organization in question, no amount of transparency will force it to change. Thus, transparency must be combined with some form of leverage. Outside of the IT world, this can take the form of withdrawing financial aid from misbehaving countries [53] or a loss of investors [30] or customers [112]. It can also be in the form of legal requirements that impose fines for violation, which may be spurred by newly revealed misbehavior [112]. In the IT world, leverage can be obtained from threats of de-peering bad ISPs [7] or distrusting misbehaving CAs, as discussed in the case of certificate transparency.

44 Weil *et al.* use the term *discloser* [112]. We use the term *information subject* as the data may also be disclosed by a third party like journalists, in which case the terminology used by Weil *et al.* breaks down.

45 See, for example, `https://locaterisk.com/`, which explicitly sells its services as a way to audit potential business partners ("Gain insight into the security situation of your business partners, such as suppliers, service providers, consultants or distributors.", quoted from the front page). Last accessed 2021-02-15.

Establishing leverage may be difficult for a scientific study, as these are usually hesitant to employ what amounts to threats (of driving off customers, pursuing legal action, or complaining to a supervisory authority). However, they can attempt to create status competition as another form of leverage, which we discuss next.

This can be difficult to do in a scientific study.

COMPETITION Exploiting competition is one of the most basic forms of leverage. It can take the form of reports comparing the CSR performance of different companies [65], or in the previously-mentioned studies that sought to reduce the amount of spam sent by corporate networks [47, 104, 105, 118]. In addition to the potential loss of public goodwill and customers, such competition may also motivating companies through peer effects / social comparison [41], a theory that predicts that status competition within a group will lead lower-ranked group members to try to improve their performance.[46]

A possible way of exerting influence is to exploit competition between companies through the use of comparisons and rankings.

In the area of privacy, competition and consumer protection are frequently considered together [61, 83], while Kerber sees the lack of competition in online markets as an explanation for the lack of privacy offerings, which he sees as a market failure [57]. This indicates that using competition to motivate changes to a privacy issue may prove difficult, however, to the best of our knowledge, it has not been attempted so far.

Existing literature sees privacy as a competitive issue, making it a promising avenue for future work.

2.5 SUMMARY

In this chapter we have considered the basic challenges of the web and email ecosystems, and how automated analyses can be conducted on them, discussing the tradeoffs between large-scale scanning and user-facing transparency tools. We have summarized a series of studies that sought to identify the determinants of effective large-scale notification campaigns that seek to inform system operators about issues detected through such analyses. Finally, we have then sought to understand how organizations make decisions in the area of security and privacy from an economic lense, and how their perspectives may be leveraged to produce desired behavior.

In this chapter, we have summarized the relevant prior literature in the field of this dissertation.

This knowledge forms the basis of the rest of this dissertation. In Part ii, we describe the design and implementation of two user-facing transparency platforms that we developed to serve as a basis for data collection and experiments. This data will then be used as a basis for a series of experiments described in Part iii, where we investigate the factors that influence operators' willingness to make changes to their systems in response to outside notifications.

We now proceed to describe our own scientific contributions.

46 We note that studies into peer effects have been criticized for methodological mistakes that have lead to self-fulfilling predictions. For further reading, see Angrist's work on the "perils of peer effects" [5].

Part II

ECOSYSTEM ANALYSIS

Promoting change requires knowing the current state of the ecosystem, and what problems it faces. Our literature review has shown that public transparency platforms can fill a gap in the current research data collection practices. We thus develop two public platforms that seek to take stock of the current state of privacy and security in two different ecosystems, and to serve as a basis for further studies.

3

THE WEB ECOSYSTEM — PRIVACYSCORE.ORG

In this chapter, we describe the design and implementation of our public transparency platform, *PrivacyScore.org*. The platform is intended as both a public tool for users, operators and regulators and as a platform for research data collection and experiments. It allows anyone to submit a domain, which will be automatically analyzed for privacy and security issues, with the results published on a website. We discuss the technology underlying the data collection and public results, as well as their limitations. Finally, we consider the ethical and legal questions surrounding a platform like PrivacyScore, and close with a discussion of the impact of the platform.

This chapter describes our transparency platform for the web ecosystem.

3.1 DATA COLLECTION

PrivacyScore collects data in two major areas: privacy and security. Scans use a combination of existing open source systems and specialized tools developed by the PrivacyScore team. They are orchestrated by a backend written using the Python-based[1] web framework Django[2] and a PostgreSQL[3] database, and run on a set of virtual machines currently hosted by the University of Bamberg.

The system collects data about privacy and security issues.

3.1.1 *Privacy*

Modern websites are highly interactive and often rely on the execution of JavaScript for core parts of their functionality. Thus, a static analysis of the source code will be insufficient to gain a realistic impression of the behavior of the website, especially when it comes to the use of external content like tracking or advertising scripts. We thus need to use dynamic analysis, i.e., actually execute the code of the website and observe its behavior in a real browser. This increases the complexity of the platform, but leads to more realistic results.

It uses a full browser for analysis to also evaluate dynamic content.

TRACKING The first version of PrivacyScore used the OpenWPM platform [38], an automated, headless Firefox browser intended for research. Due to stability issues, this has since been replaced with PrivacyScanner [87], a similar system based on a Google Chrome browser controlled via the Chrome DevTools protocol, primarily developed by Henning Pridöhl, a member of the PrivacyScore team. Both systems

It detects tracking by matching observed network requests with common block lists.

1 See https://www.python.org/, last accessed 2020-11-27.
2 See https://www.djangoproject.com/, last accessed 2020-11-27.
3 See https://www.postgresql.org/., last accessed 2020-11-27

are based on the same principle: using a real web browser to automatically access a website and log all interactions of the website (network requests, cookies, etc.). We match the generated network requests and cookies against common tracking- and advertising blocklists from the EasyList project[4] to determine if any of them belong to known trackers.

It can detect a compliance issue with a common tracking service.

NON-COMPLIANCE The automated browser also allows injecting JavaScript code into the website to collect further information. We use this feature to collect information about a privacy / compliance misconfiguration: the absence of IP anonymization for Google Analytics, a popular tracking service. At the time PrivacyScore was developed, activating this feature was mandatory for German websites that use this service. After the GDPR came into effect, using the feature is still recommended by the German data protection authorities[5].

It infers the geographic location of servers based on the IP.

SERVER LOCATION Finally, the geographic location of the web- and mailserver (if one exists) is of interest, as this has both legal implications (which jurisdiction regulates third-party access to the data stored on the machine? Which data protection rules apply?) and can influence which countries and Internet eXchange Points (IXPs) can observe traffic sent to and from the server. To determine this information, we use a GeoIP lookup, which maps IP addresses to countries. While this process isn't always perfectly accurate, it will often generate results of sufficient precision for our purposes.

3.1.2 *Security*

A set of security checks verifies if best practices are being followed.

As discussed earlier, the security of the servers in question is also of interest to the user, not to mention the system operators. We thus also conduct a series of security-related checks. While these checks do not allow us to make general statements about the overall security of a website, they can show how much effort the operators have invested in following best practices for secure web development.

Websites should use TLS / HTTPS by default.

CONNECTION SECURITY It is considered best practice to encrypt the connection from the web server to the user using TLS (HTTPS). We thus check if this encryption is offered and if the website uses it by default (i.e., forwards visitors from the non-secure version to the secure version of the website). If both versions exist, we also compare them to ensure that they show identical content, as some servers may

4 See https://easylist.to/, last accessed 2020-11-27.

5 See https://www.datenschutzkonferenz-online.de/media/dskb/20200526_beschluss_hinweise_zum_einsatz_von_google_analytics.pdf (page 6), last accessed 2021-05-17.

offer an encrypted version of the page, but only serve an error page on it.

However, even if the connection is encrypted, it may still be insecure. Over the last few years, a large number of issues were found with different TLS libraries. These vulnerabilities can often be detected through external scans. We thus use the popular *testssl.sh*[6] software to determine if the TLS configuration fulfills modern standards and all security issues are patched or mitigated. As this is natively supported by testssl.sh, we take this chance to also perform the same checks for the email server associated with the domain (as determined by its Mail eXchange (MX) record in the DNS).

The server should use mitigations for common TLS vulnerabilities.

WEBSITE SECURITY The website itself can also take measures to ensure it is more robust against compromise. To begin with, if the website is offered via an encrypted connection, the website can set the HTTP Strict Transport Security (HSTS) header[7], thereby instructing the browser to always visit the site using the encrypted version. This can protect visitors against TLS stripping attacks [74].

A number of HTTP headers can be used to improve security and privacy.

Other useful headers include the X-Frame-Options[8] header (which controls if the website can be embedded inside a frame in a different website), the X-Content-Type-Options[9] header (which disables MIME sniffing, thereby offering protection against attacks that try to fool the browser into executing external code), the outdated X-XSS-Protection[10] header (which enables certain countermeasures against Cross-Site Scripting (XSS) attacks), and the more comprehensive Content Security Policy (CSP)[11] header (which contains detailed policies about which resources can be retrieved or executed from external sources, a powerful defense against many attacks). Finally, the referrer-policy[12] header can be used to limit the amount of information the browser sends to other servers when clicking links embedded in the website. We check for the presence and values of all of these headers in the server responses collected by the automated browser used for the privacy checks.

These headers have varying levels of complexity and different use cases.

6 See https://testssl.sh/, last accessed 2020-12-04.
7 See https://developer.mozilla.org/en-US/docs/Web/HTTP/Headers/Strict-Transport-Security, last accessed 2021-03-19.
8 See https://developer.mozilla.org/en-US/docs/Web/HTTP/Headers/X-Frame-Options, last accessed 2020-12-04.
9 See https://developer.mozilla.org/en-US/docs/Web/HTTP/Headers/X-Content-Type-Options, last accessed 2020-12-04.
10 See https://developer.mozilla.org/en-US/docs/Web/HTTP/Headers/X-XSS-Protection, last accessed 2020-12-04.
11 See https://developer.mozilla.org/en-US/docs/Web/HTTP/Headers/Content-Security-Policy, last accessed 2020-12-04.
12 See https://developer.mozilla.org/en-US/docs/Web/HTTP/Headers/Referrer-Policy, last accessed 2020-12-04.

Server operators should take care not to disclose information through unlisted files.

INFORMATION LEAKS Finally, a server may unintentionally disclose information through misconfigurations or unlisted files, an issue that is more common than may be expected[13] [98]. Such files can include source code repositories, database backups, cryptographic key material, or details about the configuration or access patterns of the server. We attempt to detect such leaks using a pre-defined list of common names for unintentionally disclosed files, which we try to access. If a matching file is found, we verify if it contains potentially sensitive information using a series of heuristics like common strings that are included in database backups or cryptographic keys.

3.2 COMMUNICATING RESULTS

Results are shown on a Django-based website.

After the scanners have collected the information, it needs to be displayed to the user. Like the controller for the scanning system, the web frontend exposed by PrivacyScore is written using Django. A screenshot of the homepage is given in Figure 1. The Website can show the results in two forms: for an individual site, and for an entire list of websites.

Individual results contain information about the individual checks.

INDIVIDUAL RESULTS An individual result is shown as a report on the website, with a number of individual checks in four categories. Each check contains a brief description that is shown by default, as well as a more detailed explanation that can be displayed by clicking on the result (see Figure 2). The interface also allows the user to see when the scan was conducted, as well as a small, pixelated[14] screenshot of the website to visually verify that it was retrieved correctly by the scanning system. The user can also trigger a new scan or download the scan results in a machine-readable format.

Websites can be compared using lists.

LISTS In some cases, it may be desireable to compare multiple websites and see which one performs best in specific areas. For these cases, the system offers the option to upload a set of related websites to create a custom list. All websites in the list will then be scanned individually, and the results for the entire list will be aggregated in a dedicated, ranked report.

The ranking is based on the worst results in each of the four categories of checks.

The ranking is established by taking the worst result for each category of check (tracking, web encryption, protection against attacks, and mail encryption) as the result for the entire category. It then orders them by the results of the first category, using the second as a tie-breaker for the first, the third for the second, and the fourth for

13 See https://en.internetwache.org/dont-publicly-expose-git-or-how-we-downloaded-your-websites-sourcecode-an-analysis-of-alexas-1m-28-07-2015/, last accessed 2021-04-01.

14 We significantly reduce the resolution the screenshot to ensure that we do not inadvertently violate the copyright of the website owner by replicating the contents of the website.

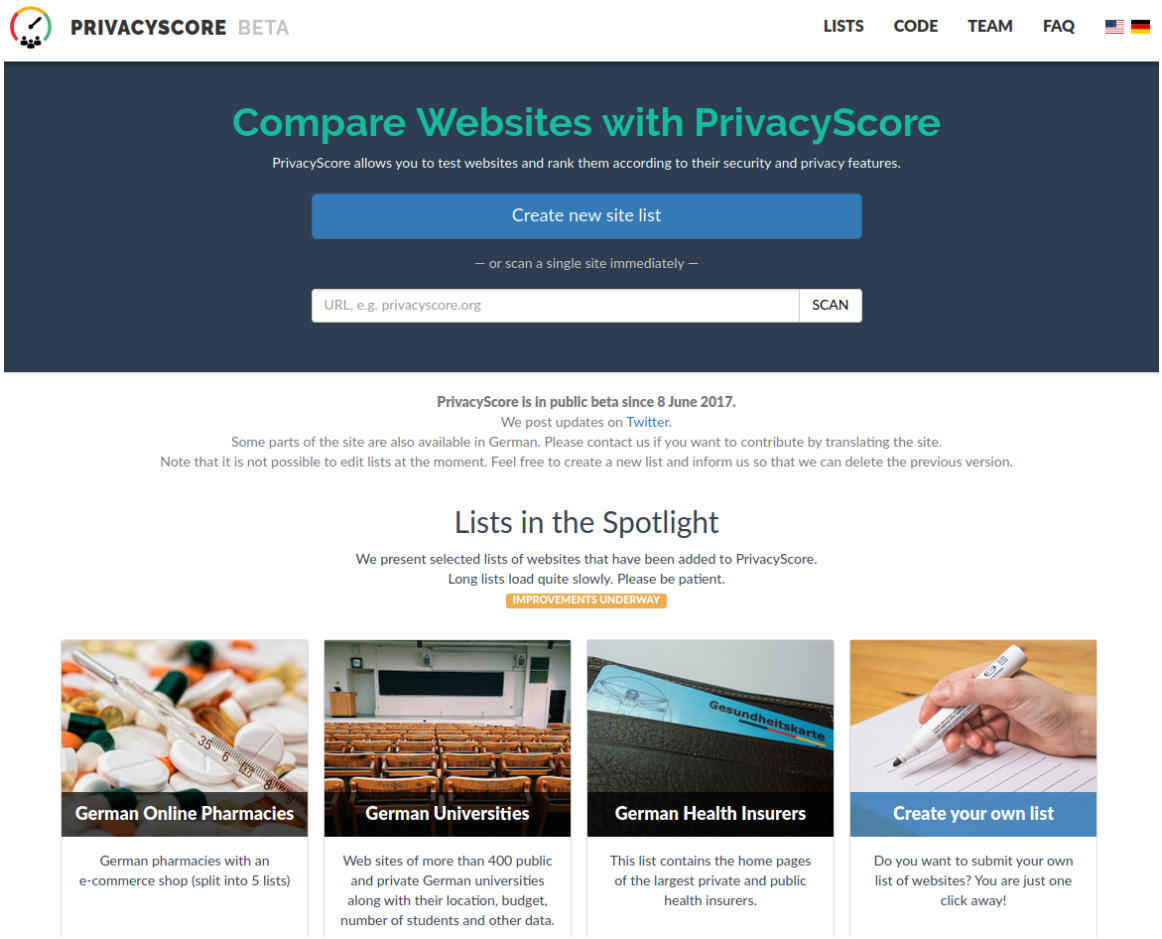


Figure 1: The PrivacyScore.org homepage showing the options to create a list or trigger a one-off scan at the top and a few example lists at the bottom.

the third. If all four categories have the same result, multiple sites can share a place in the ranking (e.g., place 3 and 4 in Figure 3). The order of the categories can be changed by the user, depending on their own priorities.

3.3 LIMITATIONS

Automated scans are imperfect.

Like every automated process involving a complex subject, an automated analysis of websites can suffer from errors. We discuss the most important error classes here.

Cookie consent scripts may block detection of third-party content.

THIRD PARTY DETECTION The most significant limitation is related to the detection of third-party content. After the GDPR came into effect, many websites started adding *cookie consent* banners. Their exact implementation varies, but many only start embedding external content once the consent has been given. Automatically detecting and confirming consent forms is a complex issue, which has not been

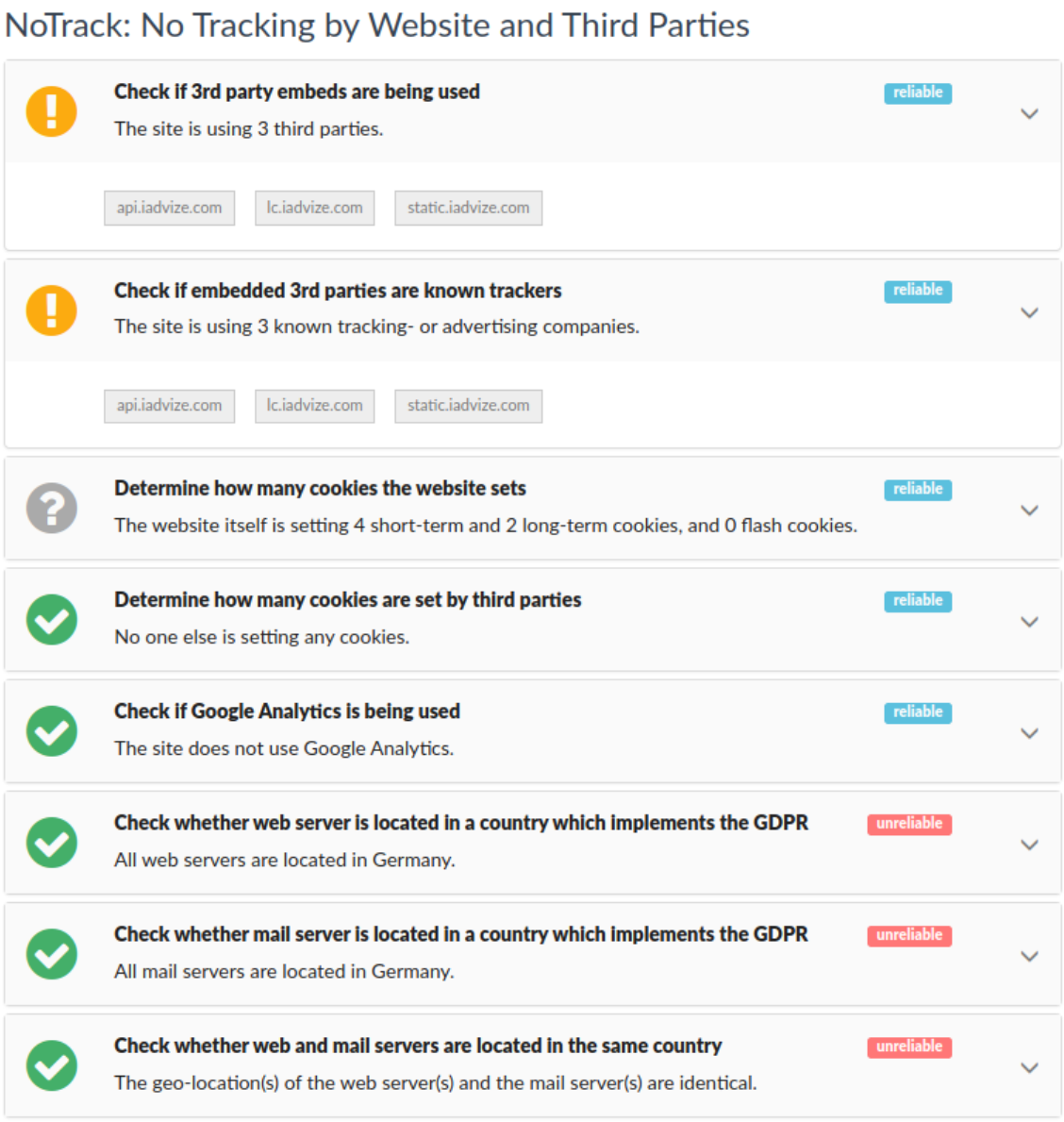


Figure 2: A detailed result for a website, showing the results for the tracking checks.

addressed in the PrivacyScore implementation so far. Thus, the system may suffer from false negatives, i.e., it may miss some third parties on some websites.

Tracking may differ between different parts of the website.

SINGLE-PAGE SCANS PrivacyScore currently only scans the page provided by the user. Vekaria *et al.* have shown that different parts of the same website may have different trackers enabled [109]. Thus, the measured privacy-friendliness of a website may depend on which exact page is added to a ranking. This could be addressed by choosing a small number of internal links from the initial page and crawling them as well. However, this would raise additional questions (e.g.,

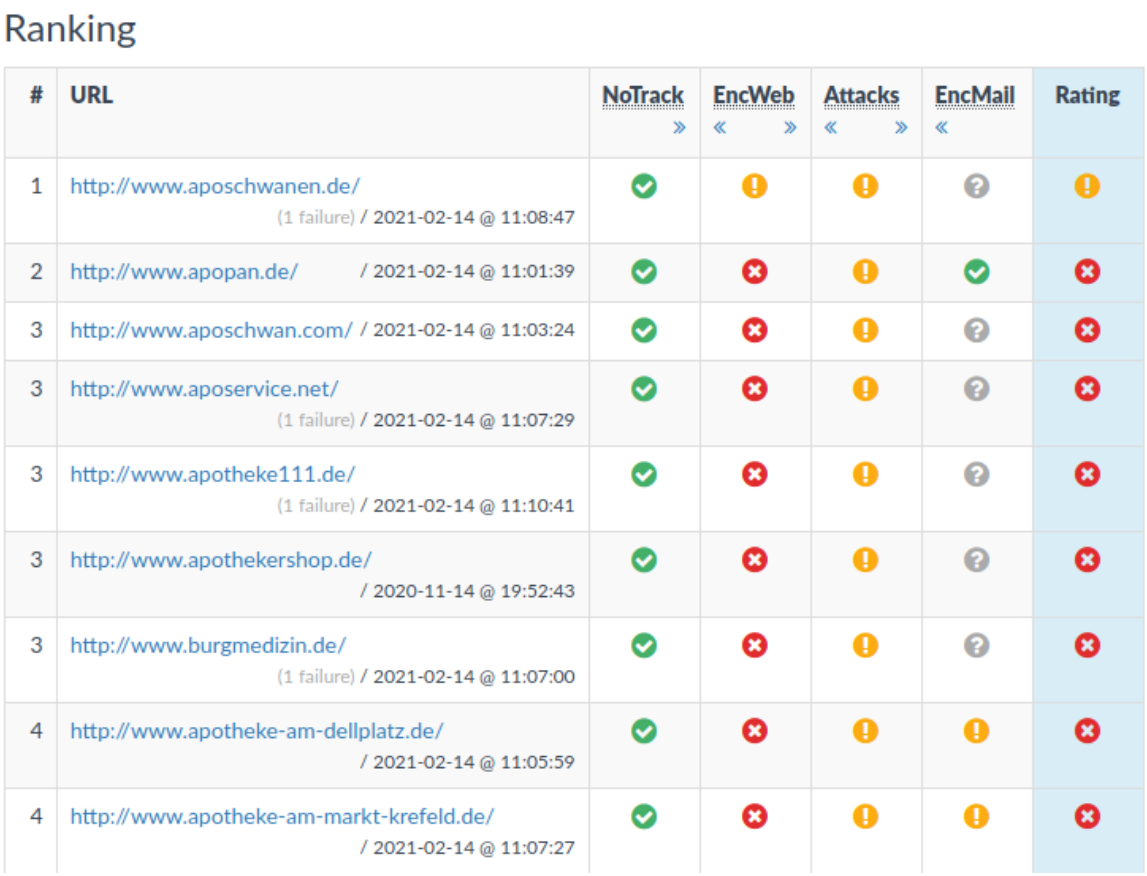


Figure 3: An example for a list of websites on PrivacyScore.

how to display the results if the pages *do* differ). Thus, such a solution has not been implemented in PrivacyScore.

IDSs and firewalls Some of the scanning techniques used by PrivacyScore may trigger automated defensive systems like Intrusion Detection Systems (IDSs). For example, email servers frequently use so-called *tarpitting*[15] to slow down spammers by reducing the speed of each subsequent connection within a short timeframe from the same IP. As a TLS scan establishes many connections, such scans frequently encounter errors or timeouts when facing such defensive mechanisms. Similarly, our checks for information leaks can also trigger IDS systems and lead to automated or manual abuse messages if they are interpreted as attacks. This also raises the question of the ethics and legality of our scans, which we discuss now.

PrivacyScore may run afoul of security tools like firewalls, leading to scan errors.

15 See https://www.techopedia.com/definition/1722/tarpitting, last accessed 2021-02-01.

3.4 ETHICAL AND LEGAL ISSUES

We consider the ethical and legal issues surrounding PrivacyScore.

Intuitively, unsolicited scans may be considered questionable from an ethical or legal perspective, especially if the results are published openly on the internet. We thus briefly discuss these two issues here.

PrivacyScore is a dual-use tool.

ETHICS OF SCANNING We aim to assist users and website operators, but not criminals, with our scans. One may consider PrivacyScore to be a dual-use tool, as it can surface security-critical information that can be abused by attackers. Given that many other public tools offer TLS scans, we consider any information derived from such scans by PrivacyScore to be unproblematic, as it can be easily obtained from other sources as well. For information leaks, the situation is less clear. We believe the scans to be acceptable, as scanning for these issues is very simple, i.e., it could be done easily by the attacker without resorting to PrivacyScore.

It also consumes resources on the scanned systems, which we address through rate limiting.

Scans also generate traffic and CPU load on the servers of the operator, which may incur costs for them. Given that they are hosting a website with the express purpose of making it available for public consumption, we believe that the resources consumed by PrivacyScore, which do not significantly exceed that of a regular page visit, are acceptable. We carefully designed PrivacyScore to avoid overloading servers by imposing a rate limit on scans — each website can only be scanned every 30 minutes. This prevents PrivacyScore from being used for DDoS attacks.

It raises several legal questions, which we address in a separate publication.

LEGALITY OF SCANNING Some site operators may also question the legality of operating PrivacyScore. Potential concerns include (intellectual) property rights, data protection, cybercrime legislation and competition law. As PrivacyScore is operated in Germany, German and EU law primarily applies. This makes it a difficult topic to discuss in an English-language dissertation. We thus refer readers to our German-language publication on this topic [71], in which we consider several potential legal issues. We briefly summarize the findings of that paper here.

Operating PrivacyScore is compliant with relevant legislation.

To determine the legality of scans, we consider the questions of data ownership, website terms of service, copyright, data protection, competition, and "hacking" laws, finding that none of them apply to the automated scans performed by PrivacyScore. We also consider the legal implications of publishing the scan results, where we find that the individual technical scan results are falsifiable statements of fact and can thus be published. Finally, we discuss if the website operators can demand that published (factually correct) results should be deleted, where we find that in most cases, the public interest in the results will outweigh the interests of the site operator, unless significant concerns stand against them (which can only be decided

on a case-by-case basis). We thus believe the operation of PrivacyScore to be compliant with all relevant legislation.

3.5 IMPACT

PrivacyScore has attracted significant traffic.

PrivacyScore has been operating in an *open beta* model since June 2017. As of May 2021, it regularly observes between 250 and 500 visitors per day, and has performed over 2.9 million scans. It has also served as the basis for research articles [82, 93, 97] and talks[16], and led to the detection and remediation of a security issue in a web shop system used by hundreds of pharmacies[17]. The technology was also used as the basis for the three studies described in Chapter 5, 6 and 7.

It also serves as the basis for three studies in this dissertation.

3.6 SUMMARY

We now proceed to describing a second transparency platform we developed.

In this chapter, we described the development and operation of the automated transparency platform PrivacyScore.org. PrivacyScore performs automated analyses of websites for privacy, security and compliance issues, and seeks to provide a useful service to end users, system operators, regulators and researchers, and serves as a technical basis for several studies, which are described in later chapters of this dissertation. However, before discussing these other studies, we first describe another transparency platform that we developed as part of this dissertation.

16 For example at the German OWASP Day 2017, the Vienna Privacy Week 2017 (both given by other members of the PrivacyScore team), and the MRMCD 2017 (given by the author of this dissertation, see https://media.ccc.de/v/DC9AG9, last accessed 2021-05-20).

17 See https://www.spiegel.de/netzwelt/web/online-apotheken-sicherheitspanne-betraf-mehr-als-170-websites-a-1209251.html. The original article on Tagesschau.de has since been deleted, an archived version is available at https://web.archive.org/web/20180524082224/https://www.tagesschau.de/inland/apotheken-datenleck-101.html. Last accessed 2021-01-14.

4

THE EMAIL ECOSYSTEM — PRIVACYMAIL.INFO

The email ecosystem has not received a lot of attention in privacy research.

The area of web security and privacy has (rightly) received a lot of attention in the past years. However, this focus on the web and, to a lesser extent, mobile applications, has led to another ecosystem being mostly ignored: the area of email tracking has only received limited attention in the past years, with only a few studies seeking to quantify its extent [37, 46, 50] or evaluate its acceptance by users [113]. No automated tools exist that allow users to determine if a newsletter is tracking them.

We present the PrivacyMail platform, which performs dynamic analyses of email newsletters.

In this chapter we present our public transparency platform for the email ecosystem, PrivacyMail.info. Our platform performs automated analyses of email newsletters and, like the previous system, reports the results on a website. In contrast to previous studies, this allows us to rely on users to perform the manual work of signing up to newsletters, a process that has proven hard to automate [37]. We describe the design and implementation of the platform as well as its limitations and the ethical and legal issues surrounding its operation. We close with a discussion of the impact of the platform in the time since its inception in 2019.

4.1 DATA COLLECTION

We describe the data collection process of PrivacyMail.

Similar to PrivacyScore, PrivacyMail is intended as a public transparency platform that makes the prevalent tracking visible to end users. In this section, we give an overview of the data collection process, from signing up a new newsletter to analyzing incoming messages. Like PrivacyScore, the platform is built using Python[1] and the Django Web framework[2].

4.1.1 *Adding a Service*

Newsletters can be registered by creating a unique email address for them.

Any service that sends out newsletters can be registered with the system by entering its URL into the system. PrivacyMail will generate a unique identity with an email address (hosted by PrivacyMail), name, and gender (as some newsletter providers ask for this upon registration), and display it to the user performing the registration. The user will then enter that email and other required information into the newsletter sign-up form. The resulting email confirmation will be received by PrivacyMail.

1 See https://www.python.org/, last accessed 2020-11-27.
2 See https://www.djangoproject.com/, last accessed 2020-11-27.

The system also collects additional metadata about newsletters.

The user will also be invited to add additional metadata about the service. This includes information about the country and industry sector of the website. This metadata can later be used for further analyses.

It will attempt to automatically click confirmation links.

Once a confirmation email is received, the system will attempt to automatically determine the correct confirmation link to click based on a number of keywords. If this is successful, the confirmation link is clicked automatically and the identity is marked as confirmed. Otherwise, the link must be found and clicked and the new identity confirmed through manual action by an administrator, and no further automated processing takes place until then. Once an identity has been marked as confirmed, any future emails from the sending domain will be automatically processed without human interaction.

4.1.2 *Analyzing Emails*

Messages are analyzed using a combination of static and dynamic analyses.

When a new email from a permitted sender for a confirmed identity arrives, it is automatically processed. First, the email is saved to the database, including all relevant headers. Next, all external links (but not the embedded external resources, like images) are extracted from the email. The system tries to detect subscription management links based on a number of common keywords, to avoid accidentally clicking an unsubscribe link. Once all likely management links have been excluded, the system randomly chooses one of the remaining links and marks it for later investigation.

EXTERNAL RESOURCE ANALYSIS A common technique for tracking email messages is to include images that are loaded from a remote server. This link is then personalized for each recipient, allowing the server operators to determine if a message was read by checking if the image was loaded. Additionally, they can inform additional trackers by forwarding the request to other tracking services which can perform their own tracking and profiling. Englehardt *et al.* observed this to be a common occurrence [37].

Email tracking frequently uses remote content like images.

The system detects tracking using an automated Firefox browser with Javascript disabled.

To detect this tracking, we save the message to an HTML file and host it on a machine-local web server. This allows us to view it with OpenWPM, an automated Firefox browser intended for research [38] that can be operated with JavaScript disabled to approximate the behavior of a mail client with remote content enabled. OpenWPM will log all requests and responses generated by viewing the email, thus giving us an accurate representation of what will happen when a user views this email without clicking any links. Using this (instead of a static analysis of embedded external content) allows us to see not only the embedded external trackers, but also any additional trackers contacted through HTTP redirects. All requests and responses and the relations between them are saved in the database.

LINK ANALYSIS In addition to tracking through remote content, newsletter providers can also track individual links by linking to their own servers, which log the visit before forwarding to the actual target of the link (i.e., the news article or product). This confirms that an email was viewed, and further contributes to the profiling of the reader by seeing which links they are interested in. Once again, trackers can also transitively forward to additional trackers before linking to the final destination.

Trackers can also use personalized links.

To detect tracking through personalized links to third party tracking services, we delete the local state (cookies, sessions, ...) of the OpenWPM browser and instruct it to visit the link we have previously selected in the email. Again, we log all requests and responses and identify the chain of HTTP redirects that takes place when visiting the link, until the final destination is reached. Any contacted domain that is not a (sub)domain of the final destination domain is interpreted as a tracker.

The system visits a random link from the message to determine which domains are contacted.

EMAIL DISCLOSURE ANALYSIS Trackers use different techniques to identify email recipients in their tracking URLs. However, identifiers derived from the email address are common. Previous work has shown that in many cases, hashes or encoded versions of the email address are used by tracking services [37, 50], in some cases nesting different encodings or hash algorithms (e.g., `md5(sha1(email))`). This shows that the email addresses of recipients are widely shared with third parties, either intentionally by the sender of the newsletter, or implicitly by the tracking services. Previous work has shown that simple hashing of such personally-identifiable information is insufficient to guarantee privacy [75].

User identifiers may be derived from the email address.

To detect this eMail leakage, we compute a series of hashes and encodings of the address, nested to a depth of 2, and check if any of them are found in any of the recorded request URLs for the eMail. If so, we assume that this request discloses the email address, and save this fact in the database. After this, processing of the email is finished.

The system detects this by looking for common transformations of the email address.

FURTHER PERSONALIZATION DETECTION Not all personalization uses identifier derived from the email address. Users may be identified by a different identifier that is linked to their identity on the server. To detect this type of personalization, we offer the option to register more than one identity per service. The system then uses a combination of email timestamps and subject lines to match newsletter messages between different identities. Once a pair has been found, the links are extracted from both and compared. If no personalization is used, the links in both messages should be identical when excluding subscription management links. Thus, if (partially) different links are detected, this is a strong indicator that they are personalized.

It identifies further personalization by using multiple identities for each newsletter.

This also allows it to detect A/B testing.

Another possibility for differing links may be the use of A/B testing, in which different versions of emails are sent out to recipients to determine which headlines are more effective at generating clicks. These practices have been observed by Englehardt *et al.* [37]. To distinguish A/B testing from other forms of personalization, we also compare the text of the messages to see how similar they are. A high similarity indicates that the same message was sent to both identities, while a low similarity indicates A/B testing.

The system tries to detect the disclosure of the email address to third parties.

EMAIL SALE AND SPAM A newsletter provider or sender may also disclose the email address to other companies to send unsolicited advertising / spam. To detect such disclosure, each recipient identity in PrivacyMail has a list of domains that are marked as legitimate senders for messages sent to this email address. If a message from a different sender arrives, we manually investigate the message to check if it is legitimate or spam. If it is legitimate, the sender is added to the list of permitted senders, and the message is evaluated normally. Otherwise, the message is marked as spam and the results for the newsletter indicate the suspicion that email addresses may be disclosed.

Additional analyses are possible using the collected dataset.

FURTHER ANALYSES Having a large archive of emails, both for a single newsletter over time and for a large, crowdsourced collection of different newsletters, will also allow us to perform additional analyses. For example, does the number of trackers increase or decrease over time? What is the influence of regulatory changes like the upcoming ePrivacy directive? For newsletters annotated with additional metadata through crowdsourcing, we can compare tracking practices between countries and industry sectors, where Haupt *et al.* found significant differences [46]. As we collect which external services are used in the different newsletters, we can also determine what the most popular third-party services are, and similarly crowdsource metadata about them (currently type of service and country of origin). Due to time constraints, we have not yet performed such analyses, but are open to sharing the dataset with other researchers.

4.2 COMMUNICATING RESULTS

The results are shown on a Django-based website.

The results for all newsletters are made available using a searchable frontend on the project website, `https://PrivacyMail.info` (see Figure 4 for a screenshot of the homepage). This allows users to check if the newsletter they are interested in has already been analyzed, and if so, which trackers it uses and to which the email addresses are disclosed.

The results for an individual newsletter contain both an overall privacy rating (see Figure 5) and a more detailed list of results that

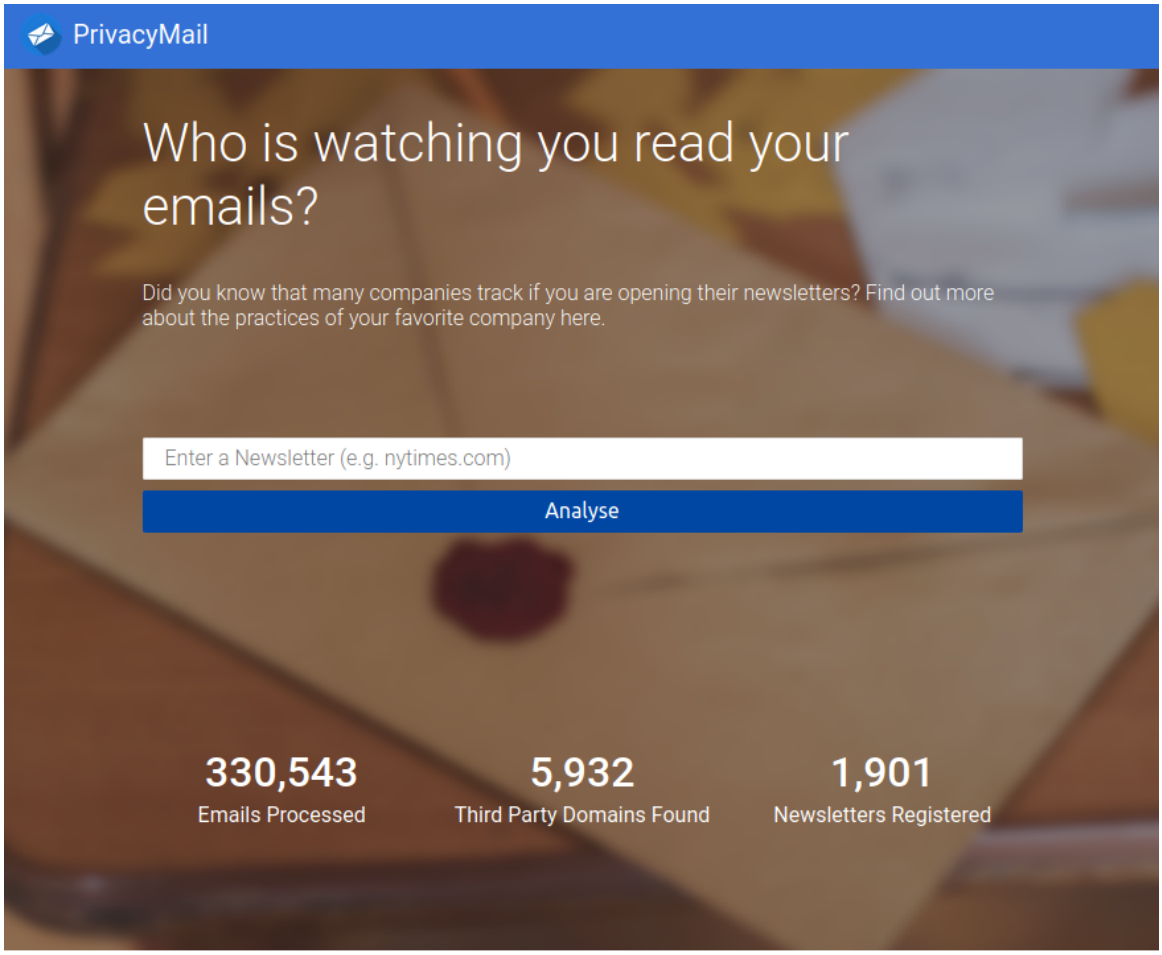


Figure 4: The homepage of PrivacyMail.info, with the option to search for a newsletter or sign up a new one on the top, and statistics about the number of processed messages below.

show which domains are contacted (see Figure 6). We crowdsource information about the embedded domains to determine if they belong to tracking companies and expose this information where it is available.

The website shows the results for individual newsletters and crowdsources further information about involved parties.

4.3 LIMITATIONS

Similar to PrivacyScore, PrivacyMail suffers from limitations, some of which we discuss below.

The architecture has some limitations.

HUMAN ERROR As we rely on untrained, non-expert users to sign up the generated email addresses to the correct newsletter, there is always a potential for error. Users may sign up for the wrong newsletter, or multiple different newsletters. We try to detect and prevent these errors using a mixture of automated and manual processes, but our mitigations may be incomplete, leading to incorrect information for some identities and newsletters.

Human error may lead to incorrect data, which we can only imperfectly mitigate.

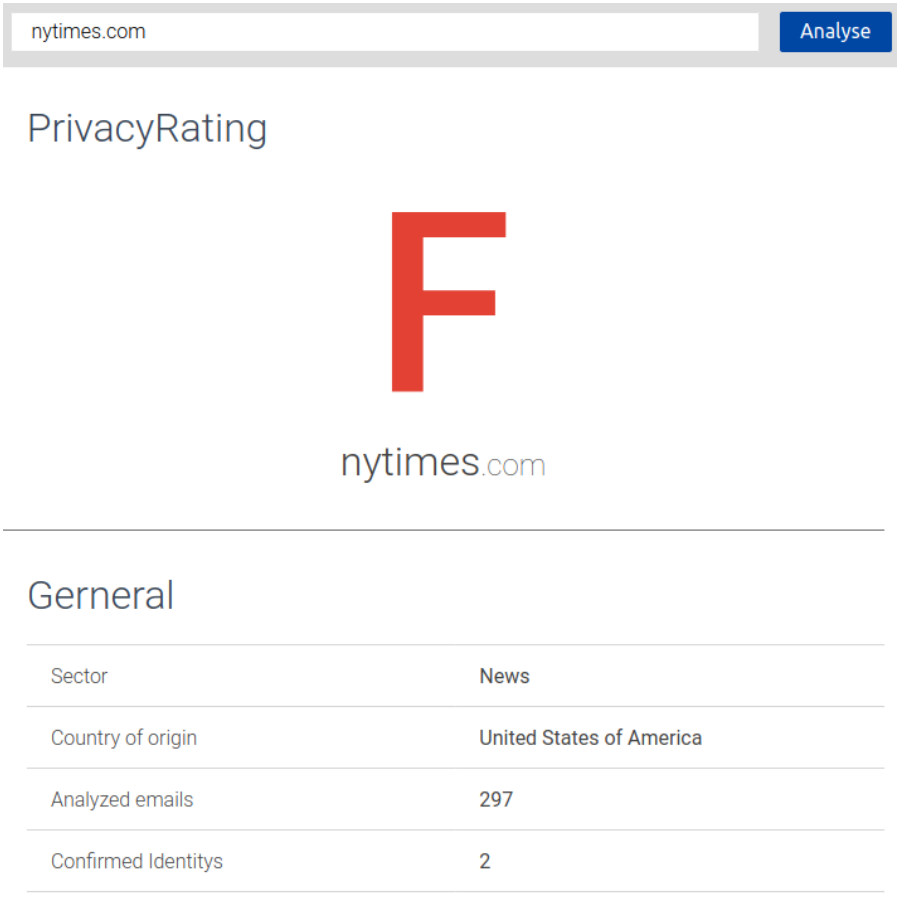


Figure 5: Summary of the rating for a single newsletter, using a US school grade system from A (best) to F (worst).

The platform is currently labor-intensive.

MANUAL EFFORT At the moment, several aspects of the system rely on manual action by the administrators to classify messages, click confirmation links, etc. We are currently working on extending the system to allow it to operate with more autonomy to reduce the overhead for the operators, for example by improving the automated detection of confirmation links.

To avoid unilateral action, we allow newsletter senders to opt-out from analysis.

MANIPULATION BY SENDERS Finally, service providers may not want their newsletters to be analyzed. As we would like to avoid unilateral action from the service providers (i.e., identifying and unsubscribing identities linked to PrivacyMail based on the used email domains), we provide them with the option to opt out of being analyzed by contacting us. To make this transparent to the users, their services will then be listed as *excluded from analysis*. However, operators may still choose to block email addresses that belong to the domains used by our service.

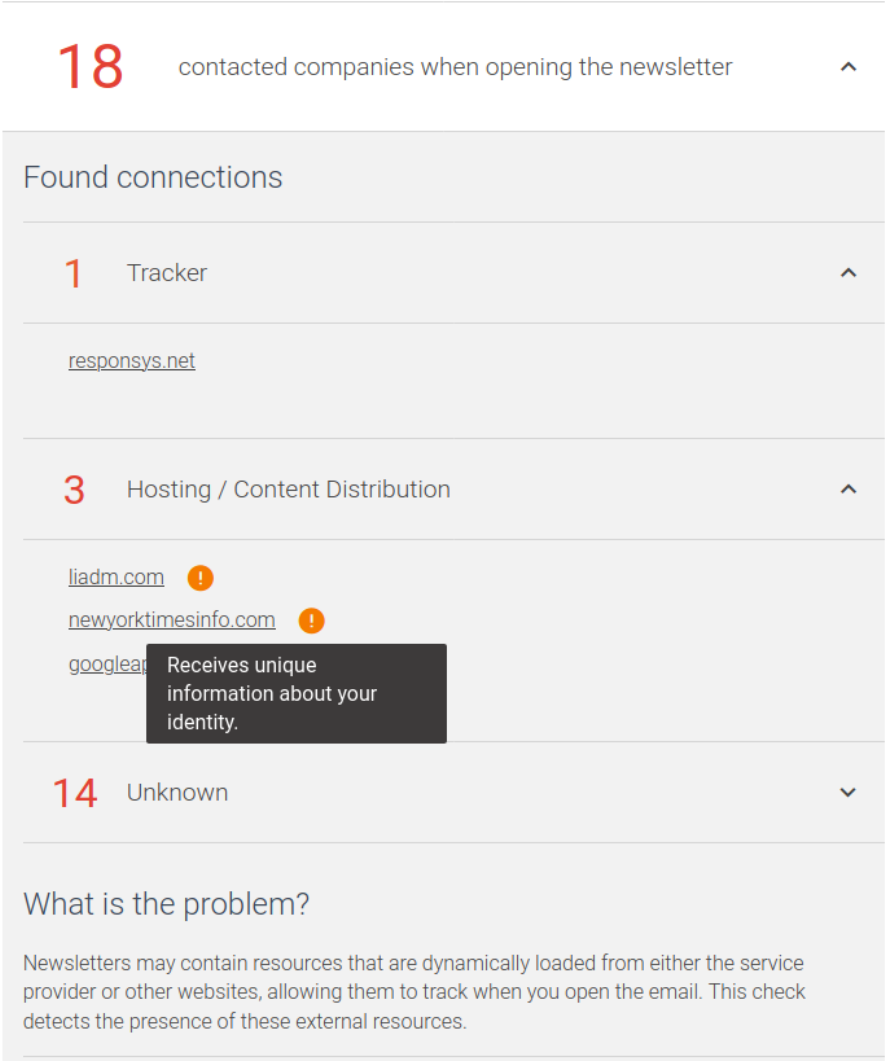


Figure 6: Detailed results for a single check. The information about the types of embedded third parties (tracker, content distribution, ...) is crowdsourced.

4.4 ETHICAL AND LEGAL ISSUES

We discuss ethical and legal issues.

Like PrivacyScore, the operation of PrivacyMail should also be considered from an ethical and legal standpoint. We thus briefly discuss these two points below.

PrivacyMail only incurs marginal costs for newsletter senders.

ETHICS OF SCANNING The first ethical question concerns the cost for the operators of the newsletters that is caused by our system. As PrivacyMail consumes even less server resources than PrivacyScore (as it does not perform complex operations like TLS scans), the pure bandwidth and computation costs imposed by PrivacyMail are minimal. Many newsletter providers charge less than a cent per recipient, making the financial costs in service fees negligible as well. The platform does not collect any information that could be used to attack the

service. Thus, there are also no concerns about dual-use tools. We thus consider PrivacyMail to be ethically unproblematic.

LEGALITY OF SCANNING While we did not perform a dedicated evaluation of the legal issues surrounding the operation of the PrivacyMail platform, we modeled it after the PrivacyScore platform, for which such an evaluation exists (cf. Section 3.4 or [71]). We thus believe these findings to also apply for this system. We do not republish the contents of the newsletter to avoid allegations of copyright infringement.

For the legal issues, we refer to the analysis of the PrivacyScore platform.

4.5 IMPACT

The PrivacyMail platform is younger than PrivacyScore, and thus had less time to establish itself. Nevertheless, it is regularly observing 50-150 unique visitors per day, and has collected a dataset of over 350 000 emails as of May 2021. In addition to the Master thesis that saw it developed, the system and its datasets have been an integral part of three further Bachelor or Master theses and two lab projects inside our group at TU Darmstadt. We also received one external data access request as part of a research project conducted at KU Leuven [54], who we supplied with the requested data. Outside of academia, the PrivacyMail platform was presented at GPN 2019[3], a conference organized by the German Chaos Computer Club (CCC), and was discussed in a podcast by a German public broadcaster[4].

PrivacyMail has a solid base of users.

It has also been used in theses and presented publicly at conferences.

4.6 SUMMARY

In this chapter, we described the development and operation of the automated transparency platform PrivacyMail.info. PrivacyMail seeks to provide a better view into the often-overlooked area of email tracking by performing automated analyses of email newsletters for different forms of tracking and publishing the results on a public website. It also collects a corpus of emails and analyses that can be used for future studies.

PrivacyMail seeks to act as a transparency platform for the email domain.

While we had originally intended to perform additional research utilizing the data collected by PrivacyMail, time constraints necessitated a focus on a different topic, which led to work on follow-up studies using the PrivacyMail dataset to be postponed. The rest of this dissertation will thus be focused on the web ecosystem, where we

Due to time constraints, we did not perform detailed analyses of the dataset.

3 See https://media.ccc.de/v/gpn19-59-analyzing-the-email-tracking-ecosystem for a recording of the talk, last accessed 2020-11-27.

4 The episode was part of the SWR Podcast "Netzagent", which appears to have been depublished on the homepage of SWR. The episode is still available on Apple Podcasts (https://podcasts.apple.com/us/podcast/id1466938159?i=1000466287752) and Spotify (https://open.spotify.com/episode/6KgjXqUJAX2GTFnYiPlp6V), last accessed 2020-11-27.

describe the results of three studies focusing on different aspects of Web privacy and security. Nevertheless, we believe PrivacyMail can prove to be a helpful tool for research into the area of email tracking, and will be happy to collaborate with other researchers, for example by providing them with datasets.

Part III

PROMOTING CHANGE

Once a set of websites is known to be affected by a privacy or security issue, we wish to encourage the operators to address it. Depending on the exact issue, different strategies may be required. We thus report on three notification studies that consider different issues and notification strategies.

5 COMPETITION

The results collected by PrivacyScore have shown us that modern websites use significant numbers of tracking and advertising services. Efforts to protect users from invasive tracking often operate through client-side blocking. However, such a solution only offers protection to a single person at a time, while changes to the website are immediately effective for every visitor. Thus, if the website operator can be convinced to voluntarily reduce the amount of tracking on their website, the impact can be significant.

Existing tracking protection mechanisms are difficult to scale.

This may seem, at first, to be an impossible task: the website operators add tracking and advertising to their website because they rely on it for their own purposes, and thus have little interest in removing it unless pressed. They gain the benefit, while the cost (in lost privacy) is borne by the visitors — it is an *externality*. However, previous work has shown that status competition based on public rankings can be an effective tool in such situations [47, 104, 105, 118].

Status competition may help to reduce tracking at the source.

To evaluate the effectiveness of such a strategy in the area of privacy and security, we leveraged the platform and publicity offered by PrivacyScore. As the subject of our study we chose a set of 152 German health insurance companies, as the importance of privacy in the area of health and medical data is beyond dispute. We created a ranking of the security and privacy aspects of their websites using the list functionality of PrivacyScore and contacted the companies via email to inform them about their results and evaluate their responses.

We utilized the PrivacyScore platform to evaluate the effect of status competition in the health insurance sector.

We begin this chapter by giving an overview of the goals and dataset of the study. We then describe the study design in more detail, before describing and discussing the results. We close with a look at the limitations of our study before concluding the chapter with a summary of the results and an outlook of future research questions posed by the study.

The resulting study is described in this chapter.

5.1 OVERVIEW

In this section, we briefly describe the questions we sought to answer and the dataset we collected for this purpose.

We first describe research questions and dataset.

5.1.1 *Research Questions*

There were two major questions underlying our study: how do website operators react when they are notified that their website has been rated in terms of privacy and security and the result has been pub-

We investigated operator reactions to external notifications with and without competition.

lished online (RQ1)? And, given previous reports about the efficacy of competitive rankings [47, 104, 105, 118], does the fact that the results are displayed as part of a ranking with their competitors change their reaction (RQ2)? To answer these questions, we considered both the changes the operators made to their websites and the responses they sent to our solicitation emails.

5.1.2 *The Issues*

We used the PrivacyScore ranking as a metric.

We generated a ranking of websites using PrivacyScore, which combines a number of different privacy issues into a single, combined ranking (cf. Section 3.2 for the details). The issues include third party tracking, cookies, the use of non-european hosting providers (where enforcement of the GDPR may prove more difficult), and other considerations. In case two websites achieve the same results in the privacy checks, other PrivacyScore tests are used as tie-breakers as per the standard ranking algorithm.

5.1.3 *The Dataset*

Our dataset contained 152 German health insurance companies.

We chose to use a dataset of health insurance companies (both private and public) for this study, as health data is an especially sensitive class of data that most people will have at least a certain degree of privacy concerns about. For this, we collected a list of German health insurance companies from the Wikipedia articles "Liste deutscher Krankenkassen"[1] and "Liste deutscher privater Krankenversicherer"[2] in the German Wikipedia, which gave us a list of 152 health insurance companies whose websites we add to a PrivacyScore list[3].

We contacted them using manually-collected contact addresses.

To contact them, we manually searched for contact information on their websites, preferring data protection contacts where possible. In total, we found data protection contacts for 96 insurance companies, which leaves 56 companies with a different point of contact, usually the address for general inquiries.

5.2 STUDY DESIGN

We used a mixed-method approach.

We use a mixed-method approach for our study [110], which has been proposed by Greenaway and Chan [44]. We combine an open question survey (to understand the perspectives of the insurance companies) with a quasi-experimental setup based on measurements by the PrivacyScore platform (to determine if the messages had any

1 See https://de.wikipedia.org/w/?title=Liste_deutscher_Krankenkassen&oldid=180861868, last accessed 2021-01-15.
2 See https://de.wikipedia.org/w/?title=Liste_deutscher_privater_Krankenversicherer&oldid=180752925, last accessed 2021-01-15.
3 See https://privacyscore.org/list/15/, last accessed 2021-01-15.

effect on the websites). We then follow a qualitative research process, which consists of planning, data gathering, preparation of analysis, analysis, and summarization [79].

5.2.1 *Experimental Factors*

To investigate the effect of competition (RQ2), we split the insurance companies into two groups, A and B. These groups received different solicitation emails. Insurers in group A received their scan results in terms of the four areas that are published on PrivacyScore: tracking and privacy, website encryption, mail encryption, and web security. Those assigned to group B additionally received the rank of their site relative to other companies in the same list, and the names of the two companies directly above and below them in the ranking. All emails also contained a brief description of the PrivacyScore project, and ended with a request to participate in the study by answering two open questions: *what do you think about such an assessment from the point of view of your company? Would you consider making changes to your website in order to improve its privacy properties?* The messages were sent from a university mail account. We give the full text of the messages in Appendix A.1.

In addition to their scan results, one experimental group was informed about the ranking, the other wasn't.

Companies were asked to comment on their results.

5.2.2 *Group Allocation*

Social comparison theory predicts that the effects of a ranking differ depending on the position of the actor in the ranking, with actors with bad rankings showing a larger effect [41]. This was also experimentally observed by Tang *et al.* [104]. We thus opted to optimize for homogeneity in rank distribution between the two experimental groups by assigning the 152 insurers to alternating groups based on their position in the ranking at the beginning of the study. This resulted in two groups with identical sizes and a similar distribution of rankings.

We used alternating group assignments based on ranking.

This process has two limitations: firstly, it led to an imbalance between private and public insurance companies: group A contains 61 public and 15 private insurance companies, while group B contains 51 and 25, respectively. Similarly, we collected a data protection contact address for 45 companies in group A (with the remaining 31 having only a general-purpose contact point), while in group B we found 52 and 24, respectively. If there is a systematic difference between these company types, it may skew the results. We will consider this in the analysis. Secondly, as we only have 152 insurance companies in our dataset and the expected response rate is low, we did not include a control group in the experimental design. This will lead to a higher number of survey responses, but means that our analysis of the effect of our messages on the websites is limited, as no unnotified group exists as a baseline.

This led to an imbalance in other factors between the groups.

We did not include a control group.

5.2.3 *Experiment Timeline*

We captured data at different times, including a long-term observation.

Before sending out the notifications, in December 2017, we saved a snapshot of the scan results for the PrivacyScore list for later comparison (termed T1). On the 8th of December, we sent out our solicitation emails to all insurers. As many insurers had not given a definitive answer after four weeks, we sent a reminder message on the 11th of January 2018. Most responses were received in January or February. At the end of February, we saved another snapshot of the scan results (T2). In March, we began to evaluate the received responses (including those sent directly to the PrivacyScore contact address) using open coding with multiple iterations. Finally, in August 2018, we saved a final snapshot of the scan results (T3) for a long-term comparison.

5.2.4 *Ethical Considerations*

Studies like this are commonplace in information systems research.

We have already discussed the ethical and legal issues surrounding the operation and use of PrivacyScore in Section 3.4. Studies similar to ours are common practice in information systems research. Our messages to the insurance companies clearly identified that they were sent as part of a research project. Answering the questions sent in the email was voluntary.

5.3 RESULTS

We now discuss the results of the study.

We now discuss the state of the websites before the messages were sent. We describe the responses of site operators, and the three types of responses we received: positive responses, complaints, and other messages. Finally, we describe how the websites changed over the course of the study timeframe.

5.3.1 *Initial State of Websites*

Before our first message, third-party tracking was common and only 2/3rds of websites had well-configured TLS.

Before we sent the solicitation messages (T1), 26 % (39) of the websites had chosen not to use any third party tracking services. 67 % (102) had a well-configured TLS setup, defined as an automated forward to the encrypted version of the website, offering TLS 1.2, and not offering SSLv2 and SSLv3. Only 26 % (39) of websites used the HSTS header. The mail server scans proved unreliable, with 6.5 % (15) of scans failing, likely due to spam protection measures like tarpitting[4]. Still, 79 % (199) of websites could be confirmed to have a well-configured TLS setup on their mail server, offering TLS 1.2 while not offering SSLv2 and SSLv3.

4 See https://www.techopedia.com/definition/1722/tarpitting, last accessed 2021-02-01.

Table 2: Recipient (Recp.) and respondent counts (#) for responses received in the two experimental groups.

Group A			Group B		
Recp.	Respondent	#	Recp.	Respondent	#
General	DPO / IT	1	General	DPO/IT	4
	Marketing	3		Marketing	1
	Board Member	1			
	Other	2			
DPO	DPO / IT	6	DPO	DPO/IT	12
	Marketing	5		Marketing	1
	Board Member	1		Board Member	2
	Other	1		Other	1

5.3.2 *Contacts and Respondents*

Many recipients had to be reminded before they responded to our questions.

Data protections contacts responded more frequently than general-purpose contacts.

The overall response rate was 27 %, with 41 insurers responding to our solicitation messages. More than half of the responses (23 out of 41) reached us only after we sent a reminder message. We contacted 97 data protection contacts (64 % of the total; A=45, B=52) and 55 general-purpose contacts (A=31, B=24). The experimental groups have similar response rates (A=26 %, B=28 %), with data protection contacts showing a higher response rate (A=29 %, B=31 %) than the general-purpose contacts (A=23 %, B=21 %). In many cases, messages were forwarded inside the contacted company and a different department responded to the message. An overview of which departments responded is shown in Table 2.

In group A, most responses came from marketing or data protection / IT teams.

The responding departments varied between the two groups. In group A, responses often came from either marketing or data protection / IT teams[5] (40 and 35 %, respectively). The remaining responses came from members of the board of directors (20 %) and other departments (5 %).

In group B, a higher proportion of responses came from data protection or IT teams, regardless of who was the initial contact.

Group B shows a distinctly different behavior. Most responses came from data protection and IT specialists (71 %), with the remainder split evenly between marketing, board members, and other departments. At first glance, this difference may be due to the higher number of data protection contact points we contacted in group B (A=45, B=52). However, the fraction of responses from data protection / IT departments is higher in group B regardless of if we contacted a data protection (A=35 %, B=75 %) or a general-purpose contact (A=14 %, B=60 %).

5 Many companies do not differentiate between their data protection and IT teams, which prevents us from making a more detailed distinction here.

Response Type	Total	Insurer Type: Public	Insurer Type: Private	Group: Group A	Group: Group B	Website improved?: yes	Website improved?: no
Positive	11 %	7 %	3 %	6 %	5 %	1 %	10 %
Neutral	11 %	9 %	2 %	5 %	5 %	0 %	11 %
Complaint	6 %	6 %	0 %	2 %	4 %	0 %	1 %*
None	73 %	52 %	21 %	37 %	36 %	11 %	66 %
Sum	100 %**	74 %	26 %	50 %	50 %	12 %	88 %

*No scan results for 8 insurers that opted to be excluded from further scans. **Deviation due to rounding error.

Figure 7: Breakdown of responses according to response type.

5.3.3 *Types of Responses*

Responses can be grouped into three categories: positive, complaints, and others.

The received responses can be grouped into three broad categories: positive responses (statements of gratefulness, expressions of interest in the study, and detailed discussions of the scan results), complaints (about unsolicited scans or the publication of the results), and other responses (acknowledgements of receipt and explicit expressions of indifference to the study).

The responses are broken down in Figure 7.

We break down the responses by their types in Figure 7. The figure also shows how the response types relate to which websites improved or did not improve over time, which will be analyzed in more detail in Section 5.3.4. We now discuss the different response types. All quotes are translations of German replies.

5.3.3.1 *Positive Responses*

Positive responses mostly came from marketing and data protection / IT departments, and varied in length.

In total, we received 16 positive responses (A=9, B=7). In group A, most came from the marketing (5) and data protection / IT departments (3), with a single response from a board member. In group B, the responses came almost exclusively from data protection or IT departments (6), with only a single response from a marketing department. Almost all respondents (A=7, B=5) reported forwarding the report to their technical staff for further analysis. The responses varied widely in length and level of detail, ranging from single-sentence responses thanking us for the information and noting that the IT department would be investigating the report further, to differentiated technical and economic analyses of the trade-offs between user privacy and economic success for their company.

Some recipients aimed to contextualize the results of the scans.

CONTEXTUALIZING THE FINDINGS Some recipients contextualized the findings of our automated scans. For example, B_{58} explained that while they use tracking services on their websites, these are disabled in sensitive areas, and noted that "*tracking is above all about the care with which data is handled. The mere collection of data does not necessarily lead to better business success.*" A_{33} noted that while they do not forward users to the secure version of their website by default, all sensitive

data (login information, payment details) are transmitted over a secure connection.

Finally, A_{91} provided a highly detailed response, going into some detail on the trade-offs between user privacy and economic success in online tracking. They noted that online tracking was required to evaluate the effectiveness of their online affiliate marketing campaigns. An additional concern is the analysis of the needs of potential customers, where they noted that "*we determine the needs of new and existing customers through market research studies. However, [...] their results are only of limited significance. This can be seen in the fact that statements determined by market research and actual user behavior partially contradict each other.*" Another concern is personalization, where the respondent argued that "*both market research and measurements show that personalized content [...] is used much more intensively by users and is increasingly expected, [which] can only be fulfilled by tracking tools and marketing automation. [...] Of course, users also have a high interest in sufficient protection of their privacy. However, we are convinced that the data protection regulations in force in Germany [...] cover the expectations of most users.*"

One highlighted the need for tracking and personalization for their company.

PROMISING CHANGES Two respondents claimed that they will be making changes based on the results of the scan: A_{67} gave a detailed response, referring to many individual test results in detail, and demonstrated that they already reacted by implementing some changes like enabling the Referrer-Policy[6] HTTP header. They also noted that some parameters related to the mail server were outside of their control, as they are managed by a third-party appliance. A_{63} responded that internal tests had confirmed the findings of PrivacyScore and mentioned nonspecific changes that were made to the website as a result. They also noted that "*the results helped us protect the privacy of the users of our website [and provided] valuable support for the implementation of changes based on an easier analysis and identification of weak points.*" In contrast, A_{35} highlighted the beta-status of PrivacyScore, but claimed that they will be performing their own checks to confirm the results, and act upon them, if necessary.

Two respondents promised to make changes.

USE OF SOFTWARE AND SERVICES Several respondents stated that they are already using external scanners to validate the security of their websites and that they previously commissioned professional security audits for their websites. On the other hand, we also saw cases of respondents not being aware of alternatives to the privacy-invasive technologies they were using, with one respondent showing surprise when being informed about privacy-enhancing alternatives (like self-hosted analytics systems) to their current practices.

While some showed a high knowledge of scanners and other tools, others were unaware of alternatives to their tracking practices.

6 See https://developer.mozilla.org/en-US/docs/Web/HTTP/Headers/Referrer-Policy, last accessed 2020-12-04.

5.3.3.2 *Complaints*

Nine public insurance companies complained to us, mostly from group B and with poor initial ranks.

We received nine complaints, all of them from public insurers. As group A contains a higher ratio of public insurers, we may thus expect more complaints to have come from this group. However, the majority of complaints came from group B (A=3, B=6). Most complaints were from companies whose site had a poor rank in the initial scan (min 25, mean 80, median 87, max 123 of 125). Four of the nine complaints (A=1, B=3) were raised directly with the contact address listed on the PrivacyScore website, i.e., not by replying to our solicitation mail. In one case, we received a positive reply to our solicitation email, while the PrivacyScore address received a complaint two days later. Eight of the nine companies requested to be excluded from future scans.

Some chose to contact the PrivacyScore team directly.

REASON FOR COMPLAINTS We observed that the tone of the responses was generally more terse in group B. Five of the six complaints from group B referenced the ranking, and one of the three complaints from group A mentioned the fact that the results are publicly available. Many were displeased that the scan was performed (B=2) and published (A=1, B=2) without asking for permission in advance. Three companies (A=1, B=2) claimed to be investigating if the scan constituted an illegal attack on their infrastructure. Moreover, B_{150} alleged a violation of competition law, and A_{109} argued that the scan violates their copyright. We had already predicted and addressed these concerns in the design phase of PrivacyScore, and our existing legal evaluation (cf. Section 3.4 or [71]) proved very helpful in responding to these allegations.

Many complaints centered on the ranking and publicity created by PrivacyScore.

Some companies (incorrectly) claimed that the scans were illegal.

AN EXAMPLE CASE Of particular interest is the complaint by B_{150}, which we discuss in more detail here. After obtaining a low rank in the initial scan, their information security officer contacted the PrivacyScore team, requesting the exclusion from future scans. Almost two months later another response reached the PrivacyScore team, this time from the chief legal officer, stating that the legal department of the company had analyzed the case and raised a number of issues with how the results were being displayed. In particular, they objected to a perceived incomprehensibility on how the rankings were computed and how the old and outdated results from excluded websites were still being displayed. They claimed that this could be "*damaging to their company*", and may be illegal under the law against unfair competition. At the same time, they acknowledged that they are open to critical analysis of their website.

We discuss one specific complaint in more detail.

It highlighted that the PrivacyScore results were hard to interprete for laypeople.

In a subsequent phone call, the company representative explicitly noted the competitive and privacy-sensitive nature of the public health insurance market and the disadvantages a company could experience from a low rank in such a privacy ranking. A constructive discussion resulted in several changes to PrivacyScore being proposed and im-

PrivacyScore was adapted to make the results easier to understand in response.

plemented, adding various clarifications to the results pages to make it easier for visitors to understand what they (do not) imply. At the same time, the insurance company started work to improve some of the security and privacy properties of its website. However, they still declined to be added back to the ranking. No legal charges were pressed.

5.3.3.3 *Other Responses*

Some companies declined to participate in the study.

We received 16 neutral reactions (A=8, B=8). Six organizations (A=4, B=2) explicitly stated that they are not interested in participating in the study. Five others (A=1, B=4) replied that our message had been forwarded internally, but never send a full response to our questions. Organization A_{153} declined participation, citing insufficient capacity, and B_{40} stated that responding would entail a work order to an external service provider. A_{51} stated that they could not give any information on the topic of privacy while B_{94} claimed that they did not understand the provided information. Finally, A_{95} simply stated that they would not be making any changes to their website.

5.3.4 *Changes to the Website*

We evaluated the state of websites before and after the messages.

To evaluate how websites change over the course of the study, we performed measurements before sending our messages (T1) and repeated them at the end of the study (T2), skipping insurers that had asked to be excluded from future scans.

Only few companies improved the privacy aspects of their websites, and the overall number of trackers increased.

PRIVACY ASPECTS We recall that twelve companies stated that they are open to changing their website (A=6, B=6). However, at T2, only 4 (A=4) of them had actually made changes to one of the parameters measured by PrivacyScore. The total number of embedded third-party trackers actually *increased* over the course of the study period, from 411 to 439. We observe 12 (A=7, B=5) cases of at least one tracker being removed, and 13 (A=7, B=6) cases of at least one tracker being added.[7] Most of these websites belong to companies that did not respond to our messages. A_{63}, who responded positively and promised to make changes, only replaced one tracker with a different one. A_{91}, who provided us with a differentiated view of the trade-offs between privacy and economic success, appears to have briefly removed a number of third parties, but has since added them again. A_{133} had stated their willingness to adapt their website — and indeed, they have removed all trackers, leaving only a cookie consent script hosted by a third party. Conversely, A_{74} responded positively to our message, but only discussed the security aspects of the PrivacyScore

7 We consider individual third parties, not the total number. Thus, if a website removes one third party and adds a different one, it will be counted in both groups.

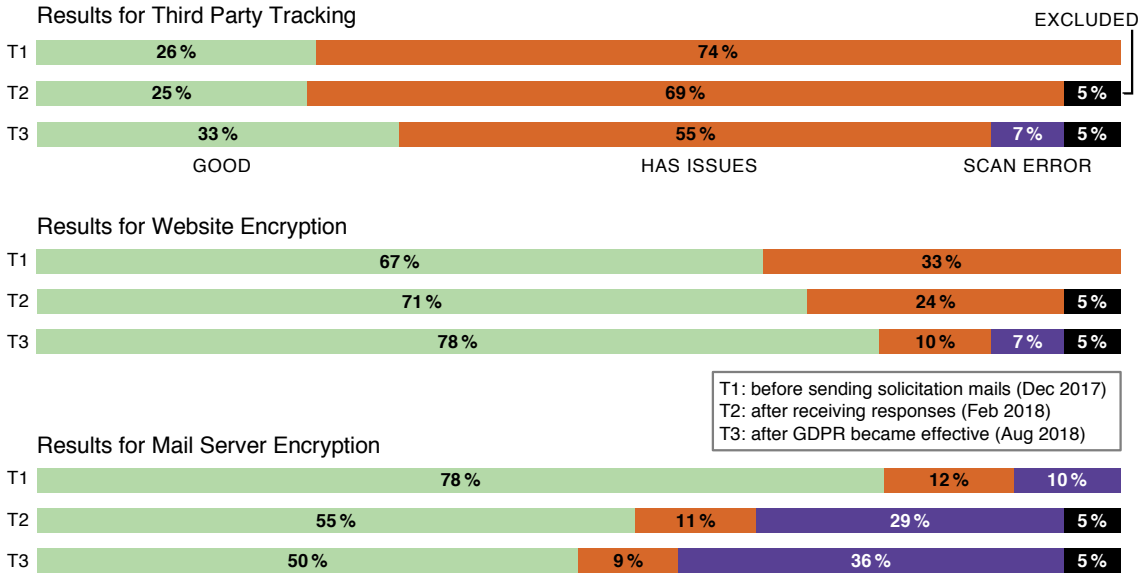


Figure 8: Scan results of health insurer websites at different points in time.

evaluation, without regard to the privacy ratings. Their website added four additional trackers over the course of the study.

The security aspects saw more improvements, although most came from companies that did not respond to our messages.

SECURITY ASPECTS Several companies also made changes to the security of their systems by changing the configuration of their TLS setups. Four companies (A=3, B=1) disabled the outdated and insecure protocol version TLS 1.0, and two of them (A=1, B=1) also disabled TLS 1.1, leaving only the latest available version TLS 1.2 active. Five companies (A=3, B=2) enabled HSTS, while one (B=1) disabled it. However, none of them responded to our message, so it is unknown if this happened as a reaction to our messages, or due to other, unrelated reasons. Similarly, 10 companies (A=6, B=4) started automatically forwarding all visitors to the secure version of their websites. Closer investigation reveals that while none of them had responded to our messages, five were maintained by the web design agency maintaining the website of B_{128}, who had asked to be excluded from future scans and threatened legal action. A manual visit of their website revealed that they also now forwarded all visitors to the secure version of their website. Thus, it is plausible that our messages are at least partially responsible for this agency-wide change.

Changes seem to be more likely if they are aligned with the interests of the company.

SUMMARY Our observations indicate that insurers are more willing to deploy changes that benefit the privacy interest of users without impacting the economic interests of the insurer (i.e., security improvements like enabling a more recent TLS versions). If privacy and company interests are in conflict, companies are more reluctant to make a change, as demonstrated by the smaller number of insurers that removed third-party trackers from their site.

5.3.4.1 *Comparison with Regulatory Changes*

We compared the observed effects with that of a regulatory change: the introduction of the GDPR.

To put our results into context we compare the impact of our solicitation mails with the effect of regulatory changes (cf. Figure 8). For this purpose, we scanned the websites once more in August 2018 (T3). Besides the introduction of the GDPR in May 2018, the timeframe also included new Payment Card Industry Data Security Standard (PCI DSS) rules coming into effect in June 2018,[8] requiring websites that process credit card payments to update their TLS configuration.

We observed an increase in failed scans, likely due to spam protection measures.

DATA QUALITY ISSUES Due to technical issues with PrivacyScore, scans for 10 sites reproducibly failed and had to be excluded in T3. Moreover, we observe significantly higher failure rates for the mail server TLS scans in T3 than in T1 and T2, precluding a meaningful comparison in this area. Cursory investigations indicate that the failures are due to the increasing prevalence of spam defense mechanisms deployed by mail server operators.

19 sites removed all third-party tracking from their websites, and others improved their TLS settings.

OBSERVED CHANGES As expected, the changes between T2 and T3 are substantial. 19 sites removed all third parties, while 4 sites reintroduced trackers (compared to 2 and 1, respectively, in time time between T1 and T2). 10 of the former websites were operated by the same association of insurers, whose websites are centrally managed. Thus, the number of distinct operators removing all trackers is at most 10. 16 improved their TLS configuration (compared to 11 in our study), with 6 adding and 1 removing HSTS (5 additions and 1 removal in our study).[9]

As expected, regulatory changes had a much higher effect than our messages.

SUMMARY Even though the time between T2 and T3 is much longer than between T1 and T2, it is unlikely that the observed differences are only due to the different durations given the changes to two regulatory frameworks. While we cannot infer what reasons served as an incentive for insurers to improve security and privacy features of their websites between T2 and T3, the combination of the passage of time and new regulatory requirements had a higher impact than our messages. This is to be expected, as the GDPR and PCI DSS changes were important and widely publicized, with strong incentives for compliance.

Nevertheless, transparency and competition may still serve as an incentive to exceed minimum requirements set out in regulation.

While regulation has a large effect, transparency may still remain a valuable tool in affecting changes, as regulatory changes are infrequent events. Furthermore, regulation only establishes a lower bound of acceptable behavior, without incentives for exceeding the minimum requirements. Since the minimum requirements still permit many

8 See https://blog.pcisecuritystandards.org/are-you-ready-for-30-june-2018-sayin-goodbye-to-ssl-early-tls, last accessed 2020-11-25.

9 The difference with the changes visible in Figure 8 is due to some websites moving to the excluded or failed category.

privacy-invasive techniques, an effective transparency regime (cf. Section 2.4.2.2) could serve as an incentive to exceed these requirements, which is vital for improving the state of online privacy.

5.4 DISCUSSION

We now interprete our results in more detail.

In this section, we discuss our results and their implications. The goal of our research was to investigate how health insurance companies react to transparency through public security and privacy ratings of their websites (RQ1) and if knowledge about the results being displayed in a ranking changes their response (RQ2).

5.4.1 *Operator Responses*

We observed a large variety in the level of detail of responses.

We find that the responses varied greatly in the level of detail and content, ranging from detailed analyses of the trade-offs between user privacy and economic success to terse legal threats seeking removal from the public ranking.

Some respondents provided detailed explanations of the rationale underlying the state of their websites.

TYPES OF RESPONSES The respondents cited a variety of reasons for the current state of their websites, ranging from conscious trade-offs between user privacy and the economic success of their company to technical limitations, e.g., third-party appliances whose configuration cannot be changed. One respondent also identified tension between the users' expectations of personalized content, which necessitates tracking, and user privacy, indicating that the company had made a conscious decision to use these technologies after considering the tradeoffs.

Some also complained, citing a fear of competitive disadvantages due to the rankings.

Public insurance companies complained more frequently.

In addition to the positive or neutral responses, we also received a number of complaints. Multiple messages cited the privacy-sensitive nature of the health insurance market, and the potential competitive disadvantages caused by a bad privacy grade, thereby implicitly confirming the premise that privacy can be a factor in competition. This led to a number of companies asking to be excluded from future scans, often under the threat of legal action based on questionable legal bases, including copyright, competition law, and cybercrime legislation. Generally, public health insurers seemed to be more likely to complain than private insurance companies.

Companies that were informed about the ranking were more likely to complain.

EFFECTS OF RANKING Overall, we observed a higher proportion of complaints in group B (A=3, B=6), which was explicitly informed that their website was part of a ranking. This indicates that the public ranking led to a higher probability of complaints being made. This explanation is supported by the fact that five of the six complaints from group B explicitly referenced the ranking.

The complaints also have in common that all of them passed through the data protection and/or IT department of the companies, either as the initial point of contact for our study, or through company-internal forwarding. Thus, at first sight, a competing explanation for the higher ratio of complaints from group B may be that the data protection and IT departments generally have a higher likelihood to complain. Coincidentally, we sent a higher fraction of our mails directly to data protection contacts in group B (A=59 %, B=68 %), which corroborates the competing explanation. However, there are two arguments that challenge its validity: firstly, several responders explicitly referenced the ranking in their complaints, and secondly, the complaint rate is still higher for group B if only messages passing through data protection and/or IT departments are considered (A=21 %, B=30 %). Thus, we give more credence to our initial explanation, which supports RQ2.

The higher rate of complaints in group B likely cannot be explained by the different distribution of points of contact in the groups.

We also note that, as discussed in Section 5.3.3.2, the complaints mostly came from websites with initially low rankings. This echoes social comparison theory [41], which predicts that members of a comparison group that are further from the top will have a stronger reaction to the ranking. It is usually assumed that this reaction is to seek to increase their position in the ranking. However, if the difference becomes too large, social comparison theory also predicts that "there will be tendencies to cease comparing oneself with those in the group who are very different from oneself" [41, Derivation D_3], and that "[t]he cessation of comparison with others is accompanied by hostility or derogation to the extent that continued comparison with those persons implies unpleasant consequences" [41, Hypothesis VI].

Our observations are consistent with social comparison theory.

PARALLELS TO CSR RESEARCH We note that some of the responses we received are similar to those observed by Lee *et al.* in the area of CSR [65]. They analyzed responses by pharmaceutical companies to a public benchmarking report by Oxfam that evaluated the companies' CSR efforts in increasing the availability of free or affordable medicine for disadvantaged populations. Among other responses specific to this topic area, the responding companies tended to express disappointment with the way they were portrayed, claimed that the report was incorrect, or tried to recontextualize their existing CSR efforts to illustrate why they perceived the rating scheme to be unfair. It seems like the same patterns also apply in the case of privacy rankings.

We also observe some similarities with prior research in the area of CSR.

5.4.2 *Changes to Websites*

Our scans indicate that the websites under study are in permanent flux, with a general trend towards increasing the number of third-party trackers. As many of the insurers whose website changed during our study did not respond to our messages, the effect of our study on the websites is hard to quantify. Based on the received responses,

Websites change frequently, and it is difficult to attribute these changes to a single source.

we can attribute four changes directly to our messages, including one company that stopped using third-party trackers altogether, and assume that at least six more changes are at least partially the result of our messages. While these changes improve the privacy and security of website visitors, the low number of overall changes shows that the effectiveness of the messages is limited, indicating that at least in its current state, transparency and rankings through PrivacyScore do not significantly influence the willingness of most website operators to change their websites. It is unclear if this is a general result or an artifact of the PrivacyScore platform, which was relatively unknown at that time, and may have thus not been able to provide sufficient embeddedness and leverage to lead to effective competition (cf. Section 2.4.2.2).

The low number of changes after our messages shows that they had only a limited effect.

This may also be because PrivacyScore was a relatively unknown platform at that time.

Many of the responses we received matched the *acquiescent approach* proposed in the IA [44] (cf. Section 2.4.1.2), stating that they are in full compliance with applicable laws without giving additional details. A small number of companies moved towards the *proactive approach* by elaborating more on their internal processes, stating that tracking is disabled in critical areas, or citing frequent security audits of their websites.

Our observations are consistent with existing theories about corporate behavior.

From our evaluation of messages and behavior, we distill three reasons for the current state of insurance company websites: conflicting value propositions, missing awareness, and negligence.

We highlight three reasons for the current state of websites.

CONFLICTING VALUE PROPOSITIONS Companies operate in a field of tension between their own economic goals (e.g., evaluating the effectiveness of their marketing campaigns), their reputation, and (sometimes conflicting) customer expectations like privacy and personalization. If not all expectations can be fulfilled, a trade-off needs to be found, leading companies to evaluate the costs of each solution. Stated in terms of the RbV [44] (cf. Section 2.4.1.2), by gathering customer data (*intellectual resource*), companies can satisfy their own goals and some of the user expectations (e.g., personalization). As the use of web tracking is ubiquitous and mostly invisible, it incurs almost no reputational cost and thus has low potential for differentiation (*relational resource*), leading most companies to pursue a *knowledge focus* and value their own economic goals higher than the privacy interests of their customers. This also explains their observed reluctance to be included in a ranking, making them favor an intentionally intransparent strategy, as proposed by Gerlach *et al.* [43].

Firstly, companies need to find a tradeoff between their economic success and the privacy of their customers. As privacy violations have few downsides for the company, choosing privacy-invasive strategies can be rational.

MISSING AWARENESS The received responses indicate that some website operators were not aware of alternative solutions that allowed them to maintain the utility of their current solutions while decreasing their impact of the privacy of their users. Such alternative solutions

Secondly, companies may be unaware of privacy-friendly ways to achieve their economic goals.

include self-hosted tracking software like Matomo[10] that keeps the data under the control of the company, or two-click social media buttons[11] that do not disclose information to social networks on every page view.

NEGLIGENCE Website operators may have been negligent and forgotten to configure their website correctly to respect the privacy of their users. This could manifest as the failure to enable the IP anonymization feature of Google Analytics (which is mandatory in Germany), or not forwarding visitors to the encrypted version of the website.

Finally, companies may simply be negligent.

SUMMARY Problems caused by negligence can in many cases be remediated by a notification to the website operators, although such notifications have been shown to not always be reliable (cf. Section 2.3). Raising awareness for privacy-preserving alternatives cannot easily be done at scale. In addition, awareness alone is not sufficient, as deciders need to be convinced that the benefits of switching to a privacy-preserving solution are worth the required effort and potential costs of changing the website, leading back to the issue of conflicting value propositions. These can only be influenced by changing the costs associated with the different options, which is easiest in the area of reputational costs. In the context of online privacy this means that privacy-invasive techniques have to become reputationally "expensive", which may turn forgoing their use into a relational resource, allowing differentiation through the *proactive approach*. However, as previously discussed, our evaluation has shown that at the moment, the effect of transparency through PrivacyScore is not sufficient for this purpose.

These three reasons can be addressed in different ways, but doing so can be challenging.

Finally, the behavior of companies can also be influenced by legislation and regulatory oversight. This is confirmed by our post-GDPR scan, which show changes in a larger number of websites. Thus, the upcoming European ePrivacy regulation is a promising avenue for affecting further changes at scale.

Regulation is a promising avenue for driving changes.

5.5 LIMITATIONS

Our study is subject to limitations: It only investigates a single, privacy-sensitive sector — health insurance — in a single country. Extending and replicating the study with different sectors and in different countries could shed additional light on the general applicability of the results. Additionally, the number of respondents and related lack of a control group do not allow us to draw statistically significant conclusions. Our analysis also only considers two fixed dates when

Our study suffers from limitations.

We only investigate a single sector and do not have a control group or statistical significance tests.

10 See https://matomo.org, last accessed 2020-11-25.
11 See https://github.com/heiseonline/shariff, last accessed 2020-11-25.

evaluating changes to the websites of the included companies, and we do not investigate if improvements are sustained over time or reverted.

The border between the experimental groups was permeable, although we received no indication that this was noticed.

Another limitation is the permeability of the group assignments. While we did not inform members of group A about the ranking feature of PrivacyScore, the scan results page contains a (non-prominently presented) link to the ranking. Thus, members of group A might have learned about it on their own. Nevertheless, we observed notable differences between group A and B, which indicates that many members of group A have not taken notice.

PrivacyScore may not have offered sufficient publicity to have an effect.

The last limitation is the nature of PrivacyScore. As a relatively new platform, the publicity provided by it may present less of an incentive for change than a more popular and well-known platform or publication would have provided.

5.6 CONCLUSION

In this chapter, we described our notification experiment with 152 German health insurance companies, where we evaluated competition as a possible factor in motivating website operators to change their websites to be more security- and privacy-friendly. Our results showed that transparency — in the form of public assessments — can improve privacy features of websites. However, such efforts can also result in complaints and legal threats. A major factor limiting the willingness to change is the conflict between user privacy and the perceived need for privacy-invasive analytics for economic success: our solicitation mails led to much larger changes in areas where company and user interests are aligned, like website connection security.

Our study highlighted the limited effectiveness of transparency provided by PrivacyScore.

Another factor contributing to the current state appears to be a lack of awareness about privacy-preserving alternatives to common tracking services. While our study provided some initial insights on these difficulties, evaluating the effects of transparency on privacy remains a promising avenue for future work, for instance when publishing assessments in more widely disseminated channels like newspaper articles. This is also demonstrated by both previous and later studies that successfully used competition to reduce the emission of spam from corporate networks [47, 104, 105, 118].

Prior and later work has shown that more widely disseminated channels can prove to be more effective.

We also learned five important lessons from this study that inform our future studies, as described in the next chapters: First, many websites may be managed by the same company, which can skew results if we do not compensate for this possibility. To ensure that the results are not dominated by changes from a single company, websites should be deduplicated based on public information (e.g., from the imprint) to at least lower the probability that a single actor can have an outsized impact on the results, as we saw in the case of a single organization making changes to 10 websites at the same time (cf. Section 5.3.4.1).

We also learned methodological lessons that inform the other studies in this dissertation.

Deduplicating based on operator instead of domain is important to avoid skewing the results.

Second, having different points of contact (i.e., general purpose vs. DPO) leads to unnecessary complexity in the analysis. It is preferable to either choose one that is used for all recipients to ensure comparable results, or to make the type of contact point an explicit experimental variable that is evaluated separately.

The point of contact should be uniform for all recipients.

Third, having an uncontacted control group is critical to be able to determine the effect of our messages. Due to the low number of insurance companies, this proved to be impossible in this study. However, this lack of comparison significantly impacts our ability to make statements about the effect(iveness) of our messages.

The sample size should be large enough to allow for the inclusion of a control group.

Fourth, measuring the impact of messages is easier if the message contains a concrete "call to action", the outcome of which can be measured automatically. An unspecific message like the one used in this study does not allow us to clearly determine how many operators took action, as there are too many possible actions they could have taken in response to viewing the report. Notifying about a specific topic like the security issues used in previous notification studies (cf. Section 2.3) makes it easier to encourage specific, measurable change.

Measuring the effectiveness of a message is easier with a more well-defined issue that can be easily detected remotely.

Finally, we saw that changes to the legal framework websites operate under can have a significant impact. This indicates that citing legal requirements may be a promising approach to promote change in future studies, at least in cases where the notified issue can be phrased in terms of data protection or cybersecurity law.

Legal requirements may prove effective in driving change.

6 ALTERNATIVE CONTACT CHANNELS

While privacy problems pit the desires of the users against those of the operator, a place where both are aligned is in the case of website security: the operator is interested in keeping their own systems secure, and the users also have an interest in the security of the website, as this reduces the chance of a data breach or malware infection, which would impact the users as well. Given this, we would assume that notifications about a security issue should be effective in driving remediation. However, prior studies have found that many different factors influence the remediation rates of notified website operators, and that the overall results can be disappointing (cf. Section 2.3). Further research may be able to shed additional light on some of these factors, like the message medium (cf. Section 2.3.1) or the framing of the problem (cf. Section 2.3.3).

Both users and operators want websites to be secure. However, this does not always translate to high remediation rates when notifying operators about security issues.

The PrivacyScore platform gives us an opportunity to conduct our own studies in this area. One of the data points collected by PrivacyScore is the presence of information leaks (cf. Section 3.1.2) — data that is unintentionally exposed by the website, like configuration information or even database backups. These issues can be surprisingly prevalent[1], and dangerous to the website [98]. Based on the preliminary data collected by PrivacyScore, we thus designed and conducted a notification study with information leaks in the focus.

We utilized data collected by PrivacyScore for our own notification study in this area.

The rest of this chapter discusses this study in more detail, starting with an overview of the goals and dataset of the study and continuing with the study design. We then go into more detail about the observed results and put them into context. After a brief discussion of the limitations of our study, we conclude the chapter with an outlook of the answers we have found, and the new questions our study raises.

We describe the resulting study in this chapter.

6.1 OVERVIEW

Here, we describe the basic research questions that guided our study, and the dataset that we collected for this purpose.

We once again begin with the research question and dataset.

6.1.1 *Research Questions*

With this study, we sought to investigate two areas of notification campaigns: the message medium and the descriptions provided in the

1 See https://en.internetwache.org/dont-publicly-expose-git-or-how-we-downloaded-your-websites-sourcecode-an-analysis-of-alexas-1m-28-07-2015/, last accessed 2021-04-01.

Our study investigated the effect of sending letters instead of emails.

notification. Most prior notification campaigns attempting to contact system operators directly relied on email (cf. Section 2.3.1). However, emails suffer from deliverability issues [17, 18, 34, 68, 98, 99] and the difficulties of automatically obtaining a functional contact address [96]. We thus investigated the use of postal letters as a contact channel, which necessitated the manual collection of contact information.[2]

We also considered different wordings of the notification as a second factor.

The second factor under investigation was the text of the notification message. Previous research indicates that more detailed messages are more successful [18, 68, 108]. We aimed to test this result by using two variants of the notification text, with one containing a more detailed description of the dangers of the reported security issues.

6.1.2 *The Issues*

We included websites suffering from five different information leaks.

For this notification campaign, we investigated unintentional information disclosure. In total, we considered five different information leaks: cryptographic keys, database backups, VCS repositories, server status pages, and PHPInfo files. We explain each of them in more detail here.

The first is the disclosure of cryptographic keys that should be kept confidential.

CRYPTOGRAPHIC KEYS Connections to servers can be secured with cryptographic protocols like TLS for web surfing or Secure Shell (SSH) for remote administration. These protocols use asymmetric keypairs to encrypt and authenticate the data. The private keys must be protected to keep the protocols secure. However, sometimes system operators accidentally place these sensitive files in public locations, where they can be found and read by third parties. In the case of TLS keys, this could allow an adversary to forge or decrypt encrypted data sent or received by the server. For SSH keys, it may allow an adversary to gain full remote access to the server to manipulate or steal data.

The second is the presence of publicly-accessible database backups.

DATABASE BACKUPS Databases frequently contain sensitive data like personal information, business transactions, or even payment information and passwords. This makes it important to secure them, both against unauthorized access and against data loss. System operators frequently create copies of their databases using tools like `mysqldump`[3], which serializes the database into a text file which can then be backed up. However, if this file is placed in a public directory on the web server, it may allow adversaries to download a copy of the database, thereby potentially compromising sensitive information.

2 While we did not compare automated and manual collection in the study, our results for manually-collected email addresses can at least be compared with previous studies. Such a comparison should be regarded with caution due to the potentially very different characteristics of the sample, but it may at least give a first indication about the effectiveness of manual collection, which can then be validated through future studies, if desired.

3 See `https://dev.mysql.com/doc/refman/8.0/en/mysqldump.html`, last accessed 2020-08-07.

VCS REPOSITORIES Software developers use VCS software like Git[4] or SVN[5] to manage the source code of their software. This source code can in some cases include sensitive configuration data like passwords or API keys, which can allow adversaries access to the system. Alternatively, the source code may contain vulnerabilities that can be found and exploited by a dedicated adversary. Thus, the source code needs to be kept secure. VCS software stores copies of the source code in hidden folders (e.g., `.git/` or `.svn/`). If these folders are made publicly accessible, this can allow adversaries to download a copy of the website source code[6]. This issue was previously considered by Stock *et al.*, who found large numbers of public VCS folders [98].

Thirdly, we consider publicly accessible VCS repositories that may contain sensitive source code.

SERVER STATUS PAGES System administrators may need to access information about their web servers to trace performance issues or application errors. Popular web servers like Apache offer special server status pages under URLs like `example.de/server-info`[7] or `example.de/server-status`[8], which contain information about used software versions or active connections to website visitors. An adversary who is able to access these pages may be able to read out sensitive information about visitors or find information about outdated and insecure software versions.

The fourth issue are public server status pages that disclose information about the server and the visitors.

PHPINFO FILES Many websites are written in the scripting language PHP, which contains a special command, `phpinfo()`[9], to print information about the PHP version, active extensions, and basic information about the Operating System (OS) of the web server. Similar to server status pages, these pages are sometimes created by system administrators looking to validate the configuration of their server. However, this information can also be helpful to an adversary that is searching for outdated and insecure software installed on a server. Under some circumstances, PHPInfo pages can even leak sensitive information directly if it is encoded in environment variables. It is thus advisable to delete PHPInfo pages once they are no longer used.

Finally, PHPInfo files can disclose information about the configuration of the web server and should not be public.

4 See https://git-scm.com, last accessed 2020-08-07.
5 See https://subversion.apache.org, last accessed 2020-08-07.
6 See https://en.internetwache.org/dont-publicly-expose-git-or-how-we-downloaded-your-websites-sourcecode-an-analysis-of-alexas-1m-28-07-2015/, last accessed 2021-04-01.
7 See https://httpd.apache.org/docs/2.4/mod/mod_info.html, last accessed 2020-08-07.
8 See https://httpd.apache.org/docs/2.4/mod/mod_status.html, last accessed 2020-08-07.
9 See https://www.php.net/manual/en/function.phpinfo.php, last accessed 2020-08-07.

6.1.3 *The Dataset*

We now describe the collection of the dataset.

To perform a notification campaign, we needed to assemble a dataset of vulnerable domains and their contact information. Here, we describe the individual steps of our dataset collection.

One part of our dataset consisted of affected German websites detected by PrivacyScore during its regular operation.

DATA SOURCES To generate our dataset, we relied on data from two sources. The first part of the dataset was based on data collected during the operation of the PrivacyScore platform (cf. Chapter 3), which searches for many of these information leaks on every scanned website. At the time of the study, PrivacyScore had found approximately 700 exposed files, spread over slightly more than 600 domains.

We focussed on German websites to avoid unquantifiable biases due to letter delivery.

Given the goal of investigating the effect of letters, we decided to limit the study to websites hosted in the country of the researchers, i.e., Germany. This allowed us to limit the financial costs of sending the letters, and also avoids the potential biases and unquantifiable effects that differences in national postal systems could have on the evaluation, for example through different message delivery times for different countries. We thus only used domains under the `.de` TLD that were still vulnerable at the start of the study. This left us with a dataset of 248 exposed files spread over 234 distinct websites.

We augmented the dataset with data from a large-scale scan of German websites based on a Wikipedia dataset.

We extended this dataset with approximately 35 000 websites collected by querying the `official website` attribute of Wikidata.org, which gives programmatic access to Wikipedia data, and again filtering based on the `.de` TLD. We used Wikidata instead of a list of popular websites like the Alexa Top Million, as the latter has documented issues with data quality [64, 94] and the research-focused Tranco toplist [64] was not yet available. Additionally, websites in the Alexa Top Million list have already received significant attention (including a set of notifications about exposed VCS files by Stock *et al.* [98]), which may introduce biases in their responses. Finally, websites listed in the Wikipedia can be assumed to have a certain level of relevance. After creating the dataset, we excluded all German universities, as they had already been part of a different study conducted by other members of the PrivacyScore team [82].

Our scanner usesd a set of static and dynamically-derived paths and verifies potential matches using regular expressions.

DETECTING EXPOSED FILES To determine which websites have exposed files, we developed a custom scanner which sends consecutive GET requests for a set of paths on each domain in the dataset and downloads the first 20 kilobytes of the response if the response code is `200 OK`, i.e. a file has been found under this path. It then determines if the detected file is an information leak based on a regular expression that matches common features of the relevant file type. Some of the requested paths are static (i.e., identical for all websites, like `example.de/key.pem` or `example.de/phpinfo.php`), while others are customized based on the domain name (e.g., `example.de/example.pem`).

The exact requested paths are documented in Table 14 in Appendix B. We ran this scan twice on different days and discarded all websites that are found to be non-vulnerable in both scans. In the end, our scans detected 1830 information leaks spread over 1736 different websites.

Vulnerabilities on different websites may have the same source. For example, two domains may be hosted on the same misconfigured server that exposes a server status page. In this case, more than one vulnerability can be remediated with a single configuration change. As this may introduce biases in the evaluation, we extracted characteristic features from each exposed file and checked if they are shared with other sites. If we detected correlated information leaks on multiple domains, we excluded all but one domain from the dataset. With this method, we were able to link 79 duplicate vulnerabilities to 23 common sources.

Where possible we attempted to deduplicate vulnerabilities based on characteristics collected from the published files.

GATHERING CONTACT INFORMATION In order to be able to send letters, we needed to manually determine a postal address for each website. German law requires website operators to provide such an address in the imprint of the website. In order to keep the results comparable, we also used manual address collection for the email addresses. We thus manually collected postal and email addresses by checking the imprints of all remaining vulnerable websites. For each medium, we searched for a technical contact and, if none was given, fell back to a general-purpose contact address, if available.[10]

We manually collected contact information from the website imprint.

The collected addresses also allowed us to further deduplicate the dataset by detecting websites run by the same operator (e.g., a publishing house or music label that operated multiple websites). This allowed us to only send a single message to the operator which notified them about all vulnerabilities at once. In the following, we thus use the term *recipient* to mean a single contact (individual or organizational) that controls one or more websites.

We used the addresses for further deduplication.

6.2 STUDY DESIGN

After collecting a corpus of websites and contact data, we proceeded to preparing and conducting the actual study, which we describe in this section.

We now describe the design of the study in more detail.

6.2.1 *Experimental Factors*

We considered two experimental factors: the message medium and the presence of a more detailed description of potential attacks enabled by the information leak. We also included a personalized link to a

Recall that we were investigating two different factors: medium and message content.

10 This decision was made before we learned the lesson about having a unified contact point from the previous chapter, and will once again cause problems in our evaluation.

self-service scanning tool (cf. Section 6.2.5) in every message, which allowed the recipients to validate if the information leak persists.

NOTIFICATION MEDIUM To investigate the role of the message medium, we compared two notification classes: postal letters (LETTER) and emails (EMAIL). Emails were sent using a purpose-specific email account linked to our research group that was only used for this purpose (`web-survey@seemoo.tu-darmstadt.de`). All emails were sent in plaintext format (instead of HTML), as Stock *et al.* previously reported HTML emails to be less effective [98]. Physical letters were using the official letterhead of our group and contained a scanned and printed signature from one of the researchers. The letterhead contained contact information for letters, email (listing the purpose-specific address mentioned above) and fax, but no telephone number. For an example letter, see Figure 16 in Appendix B.

Emails were sent from a special email account, while emails were sent from our research group.

ATTACK SCENARIOS The second factor we were interested in is the effect of providing or withholding a more detailed explanations of the risks posed by the information leak. We thus compared two different framings for our messages: in the *baseline* message, we simply gave a very short description of the issue, without going into detail as to how it could be exploited. The second class of messages, denoted with a suffix of +ATK (e. g. LETTER+ATK), additionally contained a description of an *attack* enabled by the vulnerability, under the assumption that a detailed description of the risk leads to a higher incentive to remediate.

We compared a baseline message with one that contains more detailed information about how the vulnerability can be exploited.

6.2.2 *Group Allocation*

Given these two factors, we were left with four experimental groups, plus an unnotified control group, denoted CONTROL. After the addresses were collected and before we began group assignments, we scanned all vulnerable websites again and removed those from the dataset that had already been remediated. Afterwards, we randomly assigned each recipient to one of the five groups, without consideration for address availability. Recipients assigned to a medium for which no address was available were not contacted and thus not considered in later parts of the evaluation. This slightly reduced the sample size, but avoided the introduction of a self-selection bias.

Message recipients were randomly assigned to the four groups as well as a control group.

As different vulnerabilities may show different remediation behavior, the vulnerability classes were stratified between the groups. The final assignments are shown in Table 3, which lists only recipients for which the correct address is available (i. e., those that were actually notified). As we could only send a limited number of messages, the LETTER+ATK group was slightly smaller than the other groups and contained a lower percentage of PHPInfo leaks. This leads to an imbalance, which will be discussed in Section 6.4.

Vulnerabilities were stratified between the groups.

Table 3: Number of notified recipients per group and vulnerability (recipients can be affected by more than one vulnerability).

Group	Status	VCS	DB	Key	PHPInfo	Total
EMAIL	17	18	2	1	243	275
EMAIL+ATK	13	16	3	0	253	280
LETTER	17	19	3	1	250	287
LETTER+ATK	14	18	2	0	180	213
CONTROL	21	18	4	0	269	304
Total:	82	89	15	2	1196	1359

6.2.3 *Experiment Timeline*

The experiment spanned one month.

We finalized the group assignment on June 10th, 2018, and sent the letters one day later. To compensate for the higher delivery times of postal mail, we delayed sending the emails by two days, sending them on June 13th, 2018. We then monitored all websites for a month. Due to the high cost of letters, we opted not to send reminders to unremediated sites. Instead, we ended the experiment after a month and (re-)notified all recipients (including the control group) who had not yet remediated by email, informing them that they were (still) vulnerable so that everyone had a chance to remediate the issue.

6.2.4 *Monitoring*

We monitored included sites using nightly scans.

We monitored all website that are part of the experiment using nightly scans with our automated scanning system. The results were saved to a database for later analysis.

6.2.5 *Self-Service Tool*

We offered recipients a tool to verify if the information leak had been remediated, and tracked their use of this tool.

Previous studies have repeatedly reported requests from recipients for automated tools that support their remediation efforts [17, 69, 117]. While the impact on remediation rates has been reported as low [17], a tool can still be helpful to system operators. We thus provided an online status page for each recipient where they could see the remediation status of their own website (cf. Figure 9). It showed each information leak with a short explanation of how it can be remediated and a status indicator in the form of a colored dot that is red (leak is still present), yellow (leak is no longer present, but was there within the last five days) and green (leak has not been detected for five consecutive days). The tool also allowed recipients to trigger an automated scan of their website to immediately update the

Schwachstellen-Status

Betroffene Webseite: http://[redacted].de/
Eine Schwachstelle wird als behoben betrachtet, wenn sie für mindestens fünf Tage nicht mehr festgestellt werden konnte.

http://[redacted].de/dump.sql
Status: *Behoben*

Über diese Adresse lässt sich eine Sicherung Ihrer Datenbank herunterladen. Obwohl wir deren Inhalte nicht im Einzelnen überprüft haben, lässt sich davon ausgehen, dass darin nicht für die Öffentlichkeit bestimmte Inhalte enthalten sind.

Dieses Problem können Sie ganz einfach beheben, indem sie entsprechende Datei auf dem Server in ein Verzeichnis verschieben, das nicht zu Ihrer Webseite gehört und nicht über das Internet erreichbar ist. Am besten wäre es, wenn Sie Datensicherungen gar nicht erst auf dem gleichen Server wie Ihre Webseite speichern, sondern diese auf Ihren lokalen PC herunterladen und dann auf dem Server löschen. Sie finden die Datei auf dem Server im Verzeichnis Ihrer Homepage unter dem gleichen Namen, der oben in fett angegeben ist.

http://[redacted].de/info.php
Status: *Behoben*

Über diese Adresse lassen sich Informationen über verwendete Software und deren Versionen, als auch interne Konfigurationsdetails Ihres Servers abrufen, die aus Sicherheitsgründen nicht öffentlich verfügbar sein sollten.

Dieses Problem können Sie ganz einfach beheben, indem sie entsprechende Datei von Ihrem Webserver entfernen. Sie dient nur zum Überprüfen der Installation und wird im laufenden Betrieb nicht benötigt. Sie finden die Datei auf dem Server im Verzeichnis Ihrer Homepage unter dem gleichen Namen, der oben in fett angegeben ist.

Jetzt erneut überprüfen

Das Projekt

Im Rahmen einer Bachelorarbeit haben wir tausende Webseiten auf typische Fehlkonfigurationen überprüft, die möglicherweise eine Schwachstelle darstellen. Wir benachrichtigen nun die im Impressum genannten Betreiber und untersuchen, wie diese auf die Hinweise reagieren.

Kontakt

Max Maass
SEEMOO / TU Darmstadt
Mornewegstraße 32
64293 Darmstadt

Fax +49 6151 16 - 25471
E-Mail web-survey@seemoo.tu-darmstadt.de

Datenschutz

Figure 9: The self-service tool.

remediation status after they made an attempt to remediate the issue. The page could be accessed using a personalized link included in the message, thereby allowing us to track who was using the tool, when they triggered scans, and what the results were.

6.2.6 *Evaluation*

We evaluted remediation rates over time.

For the evaluation, we considered the remediation rates over time. As previously mentioned, recipients who were assigned to a medium for which no contact information could be found were ignored in the evaluation. As not all email servers send a notice if they discard a message as spam, there may be messages in the EMAIL and EMAIL+ATK groups that were not delivered. To avoid discrepancies with the LETTER and LETTER+ATK groups, we thus did not attempt to exclude recipients with undeliverable messages from the evaluation. This lowers remediation rates compared to studies that choose to exclude bounced messages.

All vulnerabilities had to be addressed for a site / recipient to count as remediated.

We considered a website as remediated once all of its information leaks had been remediated, and a recipient as remediated once all of their websites were. This leads to all recipients making the same contribution to the overall remediation rates, regardless of the number of websites they control.

We estimated the variance of our results using bootstrapping.

Given our overall small sample sizes, we wanted to estimate how much variation we could expect on different samples with similar characteristics. For this purpose, we turned to *bootstrapping*, a method to approximate the variation of the results. This allowed us to quantify *how uncertain* we are about the results. We now briefly describe how bootstrapping works. A more complete explanation can be found in the original paper [35] or in the course materials by Orloff and Bloom[11].

AN INTRODUCTION TO BOOTSTRAPPING In the terminology of bootstrapping, the results of our experiment are called the *empirical distribution* (of (non)remediation) and denoted F^*. The empirical distribution is over our sample of n recipients, denoted $x_1, ..., x_n$, which are drawn from the *base distribution* F (i.e., the distribution we would have obtained, had we somehow been able to run our evaluation on all German websites with information leaks, instead of our smaller sample). Bootstrapping allows us to use our empirical distribution to estimate the *variation* of a statistic u computed over F by taking a sample with replacement of size n from our (known) *empirical distribution* F^*, which we denote $x^* = x_1^*, ..., x_n^*$. We can then compute the statistic of interest $u^* = u(x^*)$ over this new sample. According to the bootstrap principle [35], if we repeat this process many times, the

Bootstrapping allows us to estimate the variation of the results using repeated sampling.

11 See https://ocw.mit.edu/courses/mathematics/18-05-introduction-to-probability-and-statistics-spring-2014/, last accessed 2020-11-10.

variation of the obtained u^* approximates the variation of u. We can thus compute many resamples x^*, compute u^* from them, and then compute the measure of choice for the variation of u^*.

LIMITATIONS OF BOOTSTRAPPING The bootstrapping technique has some limitations. In particular, it assumes that $x_1, ..., x_n$ are independent (i. e., that the result of x_a does not in some way depend on the outcome of x_b, with $a \neq b$), and that the empirical distribution F^* is a representative sample of F. We tried to ensure the former by considering recipients instead of websites (thus sidestepping the issue of correlated remediations for multiple websites run by the same operator), although additional connections between websites could exist that we were not aware of (like being run by the same web design agency). The latter issue hinges on the representativeness of our sample of websites, and what we consider to be our base distribution for which it should be representative (all websites? German websites? German websites fulfilling the Wikipedia relevance criteria?). Thus, the results should be interpreted with care. However, bootstrapping still gives us more solid results overall.

Bootstrapping has some limitations, which we tried to address in our methodology.

For our evaluation, we used bootstrapping with 10 000 iterations to compute the 1st and 3rd quartiles (denoted Q1 / Q3) in addition to the median of remediation rates on each day.

We ran 10 000 iterations of the bootstrapping algorithm.

6.2.7 *Ethical Considerations*

Whenever we perform large-scale vulnerability scanning, we operate in a gray area. We have previously discussed the legality of such scans (cf. Section 3.4), and the results also apply here. To avoid collecting sensitive data, we limited the amount of data we retrieve with our scans to 20 kilobytes. This allows us to avoid accidentally downloading entire database backups and also reduces the impact we have on the system under test. All downloaded data was deleted after the end of the study.

We took care to ensure the legality and ethics of our research.

Our scanning system identified itself with a custom user-agent string and a reverse DNS entry for the IP of the scanning machine. It hosted a small website with an explanation, thereby allowing system operators to easily determine who is scanning them and why. It also offered them a way to opt out of the study.

We did not attempt to hide the source of the scans.

Our notification messages clearly stated that they were sent as part of a study and contained our contact information to allow recipients to opt out of the study. Recipients who were not notified during the study (due to a lack of available contact information with the assigned channel or because they were assigned to the control group) were notified after the end of the study to give them an opportunity to remediate and also the option to opt out of the study. We received no opt-out requests.

Messages clearly stated that they were sent as part of a study.

Table 4: Reachability of the recipients per contact group.

Group	Assigned	No Contact		Bounced		Reached		Unknown	
EMAIL	302	27	(8.94 %)	4	(1.32 %)	76	(25.17 %)	195	(64.60 %)
EMAIL+ATK	302	22	(7.28 %)	6	(1.99 %)	74	(24.50 %)	200	(66.23 %)
LETTER	304	17	(5.61 %)	3	(0.99 %)	97	(32.01 %)	187	(61.51 %)
LETTER+ATK	224	11	(4.89 %)	3	(1.34 %)	58	(25.78 %)	152	(67.86 %)
Sum	1132	77	(6.80 %)	16	(1.41 %)	305	(26.94 %)	734	(64.84 %)

We did not seek ethics approval for this study.

At the time of the study, TU Darmstadt did not require ethics approval for this type of research. We thus did not apply for ethics approval. However, we later sought out and received ethics approval for a substantially similar study employing similar safeguards, which will be discussed in the next chapter.

6.3 RESULTS

We now present the results of our study.

We now investigate the effect of our notifications, the use of our self-service tool, and briefly discuss the interactions with the recipients. Once again, the interpretation of the results will follow in the next section.

6.3.1 *Remediation Rates*

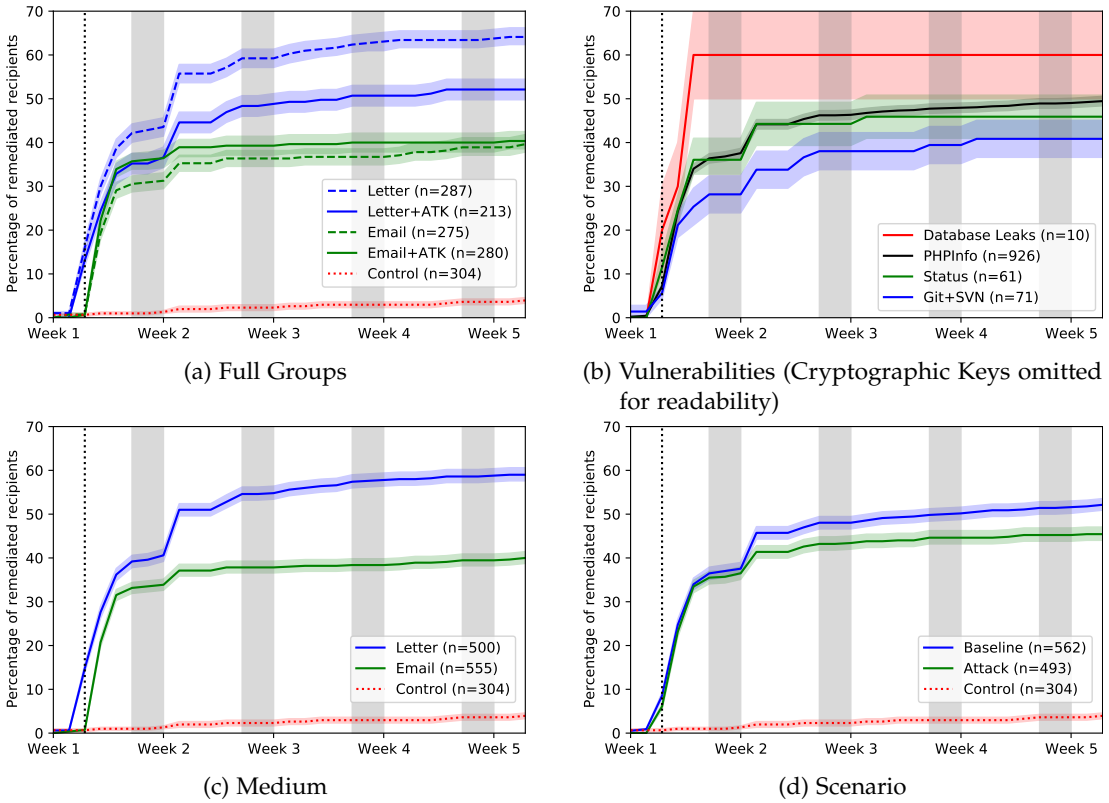
Delivery success was reasonably high.

The overall delivery success is reasonably high. 77 of 1132 recipients (6.8 %) were assigned to a medium for which no contact information was available, and thus not notified. An additional 10 emails (1.7 %) and 6 letters (1.1 %) were returned as undeliverable. As previously mentioned, the real number may be higher for emails, as not all mail servers notify the sender about rejected or dropped messages. 305 recipients (26.9 %) either sent a non-automated response to the message or used the self-service tool. We thus consider them *reached* for the purpose of the evaluation. The remaining 734 messages (64.8 %) are in an unknown state. Table 4 provides an overview of these numbers, broken down by experimental group.

All groups outperformed the control group.

Remediation rates for notified groups ranged from 39.3 to 64.3 %.

OVERALL REMEDIATION We observed large differences between the different experimental groups. The least effective group, EMAIL (without attack scenarios) achieved a remediation rate of 39.3 %, while the best group, LETTER, achieved 64.3 % remediation, an improvement of 25 percentage points. However, even the worst group outperformed the remediation rate of the CONTROL group, which only achieved 4.3 %.



(a) Full Groups

(b) Vulnerabilities (Cryptographic Keys omitted for readability)

(c) Medium

(d) Scenario

Figure 10: Median, 1st and 3rd quartiles for the remediation rates in different experimental groups, vulnerabilities, mediums, and presence of attack scenarios. Gray areas denote weekends, dotted line shows when the emails were sent.

Figure 10 shows that most remediations occur within the first two weeks of the study. We summarize the results in more detail in Table 5.

Letters outperformed emails by 19 percentage points in aggregate.

COMMUNICATION MEDIUM Letters resulted in a substantially higher remediation rate than emails (59 vs. 40 % remediation, cf. Figure 10(c)). Their effect was spread over a longer timeframe — while the email groups did not show large jumps in remediation after the first week, letters still increased remediation in the second week and only fall off afterwards.

Adding the attack details had inconsistent results between the groups.

ATTACK SCENARIOS Curiously, the impact of adding a description of attack scenarios differed between the two communication mediums: For emails, the addition of attack scenarios led only to a barely noticeable increase in remediation rates, while for letters, attack scenarios actually *decreased* the the remediation rates by 12.4 percentage points (cf. Figure 10(a)). In aggregate, this led to messages that include the attack scenario achieving a remediation rate of 45.3 % and thereby being outperformed by the 52 % remediation rate of their less explicit counterparts.

Table 5: Median, 1st (Q1) and 3rd (Q3) quartiles of bootstrapped remediation rates in percent for different groups at the end of the study timeframe.

Group	n	Median	Q1	Q3
EMAIL	275	39.3	37.5	41.5
EMAIL+ATK	280	40.4	38.2	42.1
LETTER	287	64.3	62.6	66.1
LETTER+ATK	213	51.9	49.5	54.2
All emails	555	40.0	38.4	41.3
All letters	500	59.0	57.4	60.4
All baseline	562	52.0	50.8	53.5
All +ATK	493	45.3	43.7	46.8
PHPInfo	926	48.3	48.3	50.5
VCS	71	40.8	36.6	45.1
Status	61	45.9	41.0	50.8
Database	10	60.0	50.0	70.0
Keyfile	2	50.0	50.0	50.0
CONTROL	304	4.3	3.6	4.9

We investigate this counterintuitive result in more detail by considering the different vulnerability types separately. Table 6 shows the remediation rates for the three largest groups of vulnerabilities separately for the different groups. It shows that the reduction seems to be dominated by the PHPInfo group, whose remediation rate decreased by 8.3 percentage points when including an attack scenario. This decrease is not spread evenly between the contact mediums: the EMAIL and EMAIL+ATK groups have almost identical performance (40.7 vs. 41.1 %), while LETTER+ATK showed a large decrease compared to the baseline LETTER group, lowering the remediation rate from 65.2 to 51.1 %. The other vulnerabilities showed increases in remediation of 6.3 to 10.7 percentage points when adding attack scenarios. However, the latter numbers should be treated with caution due to the limited number of samples, which lead to a large spread in the quartiles of the bootstrapped distributions, decreasing the confidence in their accuracy.

This effect was dominated by the PHPInfo group, other groups saw increased remediation rates when the attack details were present.

VULNERABILITY TYPE The different vulnerability types show small differences in their remediation trends. PHPInfo, server status and VCS leaks show remediation rates of 48.3, 45.9 and 40.8 %, respectively. The two most severe vulnerabilities, public database backups and

The different vulnerabilities showed marginally different remediation rates.

Table 6: Median and quartiles of bootstrapped remediation rates for different vulnerability types at the end of the study timeframe, with and without attack scenarios. (Database and keyfile omitted due to low sample size).

	Group	n	Median	Q1	Q3
Baseline	PHPInfo	493	53.3	51.9	55.0
	Status	34	41.2	35.3	47.1
	VCS	37	37.8	32.4	43.2
Attack	PHPInfo	433	45.0	43.4	46.7
	Status	27	51.9	44.4	59.3
	VCS	34	44.1	38.2	50.0

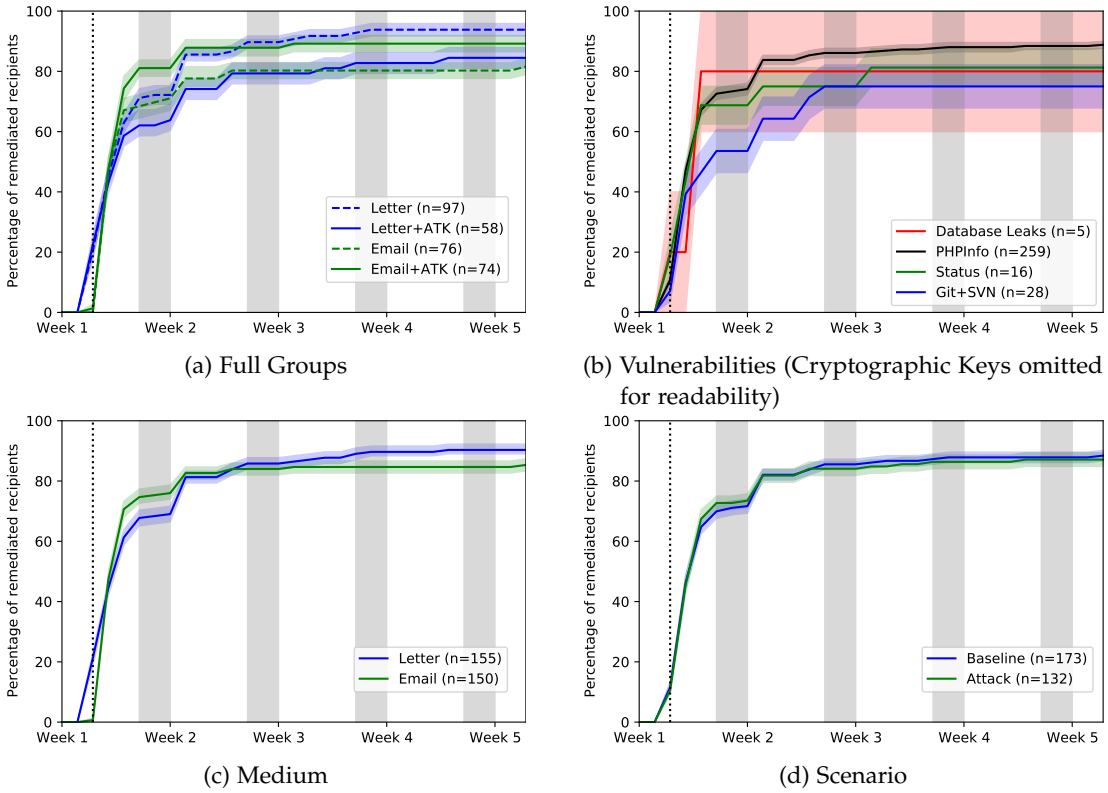
cryptographic keys, show remediation rates of 60 and 50 %, respectively. However, their small sample sizes of 10 and 2 lead to limited expressiveness.

EFFECT OF REACHABILITY Some of these results, in particular those for the message medium, may also be explainable through general differences in message delivery success: It may be that we are only measuring if letters are *delivered* more successfully than emails, instead of measuring if a letter is more effective than an email if both have been received and read. One could argue that this difference does not matter (or is even desireable), as we are primarily interested in the probability of a single sent (not: read) message causing a recipient to remediate. However, we can still try to begin to answer the second question by considering only those recipients that have either sent a manual response to our message or used our self-service tool. This (highly self-selected) subsample contains only those recipients where we can be certain that they read and considered the message — if we still observe differences here, we may reasonably expect them to be caused by the contents and medium of the message, not the success of the delivery.

The higher remediation rates for letters may have simply reflected better delivery rates than for emails.

In this subsample of 305 recipients, the remediation rates were (as can be expected) very high: 85.3 % for emails and 90.3 % for letters (cf. Figure 11). Considered in isolation, the baseline and attack variants of the message resulted in very similar performance (88.4 and 87.1 %), but when taking them together with the medium, we see that the attack scenario still helped for emails and hurt the letters, although the difference is less pronounced now. The trends for the different vulnerability classes remained similar to the full dataset as well.

However, even when considering only recipients that were definitely reached, letters outperformed emails.



(a) Full Groups

(b) Vulnerabilities (Cryptographic Keys omitted for readability)

(c) Medium

(d) Scenario

Figure 11: Median, 1st and 3rd quartiles for the remediation rates in different experimental groups, vulnerabilities, mediums, and presence of attack scenarios, for reached recipients only.

6.3.2 *Use of Self-Service Tool*

The self-service tool was widely used, with 266 recipients (25.2 %) accessing the tool at least once, and 192 of them (18.2 % of the total, 72.1 % of tool users) triggering a manual scan. The spread between the different experimental groups is given in Table 7. Many status pages (65.8 %) were only accessed on a single day, while a small subset was accessed on up to 15 separate days (median: 1, Q1: 1, Q3: 2). The triggered scans show a similar behavior: 37 % of tool users required only a single scan, while some used up to nine scans (median: 2, Q1: 1, Q3: 3). Surprisingly, only 116 recipients (60.4 % of scan users) scanned their website after remediating to validate that they were successful.

The self-service tool was widely used, though only 60.4 % of users verified their successful remediation attempt with it.

Tool users were very likely to remediate. 90.2 % of recipients that accessed the tool remediated, and this percentage climbs to 95.8 % when considering only those that triggered at least one manual scan before attempting their remediation. We stress that this does not imply that the tool *causes* remediation, as tool users are self-selected: they received and read the notification and trusted it enough to click a link from an unknown sender. This makes them more likely to remediate, with or without a tool. Thus, we cannot make any claims about the effect of providing a tool on remediation rates.

The (highly self-selected) sample of tool users was very likely to remediate.

Table 7: Percentage of recipients who viewed and used the tool before (B) and after (A) remediation.

Group	$View_B$	$View_A$	Use_B	Use_A
EMAIL	23.6 %	13.5 %	13.5 %	7.3 %
EMAIL+ATK	22.1 %	15.4 %	15.0 %	11.4 %
LETTER	25.2 %	22.0 %	18.5 %	15.0 %
LETTER+ATK	22.0 %	15.4 %	15.9 %	9.8 %

Table 8: Number of contacted recipients and non-automated responses by group and medium.

Group	n	Email	Phone	Fax	Letter	Sum
EMAIL	275	29	3	0	0	31
EMAIL+ATK	280	28	2	0	0	30
LETTER	287	25	3	1	0	29
LETTER+ATK	213	13	1	0	1	15
All	1055	95	9	1	1	105

6.3.3 *Communication with Recipients*

10 % of recipients responded to the notification message.

The overall response rate (not counting bounces and autoreplies) was 10 % (105 out of 1055). Most responses expressed gratitude for the notification, although two recipients interpreted our messages as fraud or spam and complained. To the latter, we clarified our intentions and offered not to contact them again. 86 respondents informed us that the problem had been addressed or forwarded to the maintainers of the website, although three of these still had unremediated PHPInfo leaks at the end of the study timeframe. Some recipients explicitly commented on the tool, calling it helpful, and some asked if we could scan other websites under their control. These, we referred to the PrivacyScore platform. We did not receive any opt-out requests, although our network operator received one abuse message about our scans which did not contain enough information to assign it to a domain.

Most were grateful or requested help, only few were sceptical or hostile.

Most responses were sent via email, although we also received a number of phone calls.

95 responses used email as their medium of choice, regardless of the medium we used to contact them (cf. Table 8). We also received one letter and one fax and, most interestingly, nine phone calls. The latter apparently searched out a phone number from the website of the research group, with at least one reaching the secretary of the group, and another one calling the central switchboard of the university and being forwarded over multiple intermediaries until they reached the correct person. The number of phone calls should

thus be considered a lower bound. Many of these callers wanted to validate the authenticity of the message on a different communication channel, as they mistrusted it and suspected a forgery.

6.4 DISCUSSION

Notifications were effective at promoting remediation, but the performance varied between the groups.

The fact that 48.9 % of notified recipients remediated, compared to 4.3 % of the control group, shows that our notifications were effective at increasing the remediation rates. However, we also observed a large spread of remediation rates between the different experimental groups, which ranged from 39.3 to 64.3 %. This shows that the varied factors can have a large impact on the remediation rates. We thus consider these factors in the following.

Many previous studies struggled with finding a valid contact address for their notifications.

MANUAL ADDRESS COLLECTION IMPROVES DELIVERABILITY Many previous studies attempted to reach website operators directly [17, 18, 34, 68, 69, 98, 99, 108, 117] (instead of intermediaries like hosting providers), and frequently reported delivery issues when using fully automated address collection. For example, contacting email addresses collected from the WHOIS interface repeatedly led to bounce rates exceeding 10 % [98, 99], although Zeng *et al.* also reported rates as low as 3 % [117]. When using auto-generated email addresses following RFC 2142 [24] (abuse@domain.tld, ...) the bounce rates frequently exceed 50 % [18, 96, 99]. This demonstrates that automated address collection or generation struggles to reliably reach a large fraction of recipients, especially now that the WHOIS system no longer lists email addresses for many TLDs.

Manual address collection has not received significant attention so far.

Stock *et al.* attempted to circumvent this issue through manual address collection, similar to the experiments performed in this study but at a smaller scale (N = 364 over 10 groups) [98]. They reported no bounces from their emails, but a 26.8 % bounce rate for letters. While their experiments suffer from the small sample size and self-selection (manual notifications were only attempted with websites that had not reacted to an automated notification), they nevertheless indicate that the delivery success of manual address collection may not always be as high as in our study, where we observed bounce rates of 1.1 and 1.7 % for emails and letters, respectively.

The work involved in it may be justified for very important notifications where reliable delivery is paramount.

It is clear that manual address collection does not scale to very large notification campaigns. However, it may have a place for smaller or very important notifications. Overall, a reliable, automated method for reaching system operators is urgently needed to facilitate the distribution of security notifications.

Letters outperformed emails, likely due to a combination of a priori trust and better deliverability.

LETTERS ARE EFFECTIVE In our experiments, letters substantially outperformed emails in terms of remediation rates, showing an increase of almost 20 percentage point. This is likely related to a com-

bination of factors. It is possible that letters simply receive a higher *a priori* trust than emails, as the amount of spam and scam messages sent via postal mail is much lower than that for emails. It may also be due to the lack of automated spam filters on postal messages. However, letters lead to an improvement in remediation rate even when we only consider recipients that definitely received and read the message. Thus, improved delivery alone cannot explain the improved remediation rates achieved with letters, and other factors must be at play as well.

However, it is expensive and labor-intensive, so it may not work for all scenarios.

Regardless of the source, this improvement comes at a cost: even leaving aside the issue of manual address collection, sending 500 letters cost around 400 € in postage and five hours of work to print and place the letters in envelopes. This investment may be justified for very important notifications, but is likely unsuitable as the default mechanism for notifications, even taking into account commercial mailing services and enveloping machines that could be used to reduce the required manual effort.

VERIFIABILITY FOSTERS TRUST The first step in achieving remediation is to gain the trust of the recipient and convince them that the message is authentic. Several recipients mentioned that they recognized the name of the university, which helped them overcome their initial distrust. However, we also encountered some distrusting recipients that went to considerable effort to validate the authenticity of the message through phone calls, sometimes initially reaching the wrong person. This is in line with prior results that reported recipients reaching out to verify the veracity of the notification message [15, 16].

Recipient trust is an important factor in notifications. Some recipients reported recognizing the university as trust-promoting.

Another factor that reportedly increased trust was the self-service scan tool we offered. While prior work indicates that tools do not significantly increase remediation rates [17], a tool could nevertheless increase trust and decrease support requests by recipients (if it contains sufficient documentation for the recipient to remediate the issue). Such tools have been repeatedly requested in previous studies [17, 69, 117].

Providing a self-service scan tool also increased the perceived trustworthiness for some recipients.

TANGIBLE EXPLANATIONS HAVE UNCLEAR IMPACT Previous research has shown several times that more detailed messages increase remediation rates [18, 68, 108] and trust in messages [98]. We were thus surprised that the attack scenarios did not substantially improve remediation for emails, and actively *detracted* from it for letters, at least for the PHPInfo vulnerability.

Despite the results of prior research, detailed attack scenarios did not uniformly increase remediation rates.

A different explanation holds that an expanded explanation provides a more nuanced view of the danger, which may reduce the perceived danger compared to categorical statements like "this file should not be available". This may explain why the PHPInfo sites were the only group where the attack scenario did not improve remediation. The recipients may have also done their own research: an

It may be related to the added nuance a more detailed explanation provided, or to other, unknown factors.

online search for "phpinfo dangerous" returns some results that imply that exposing PHPInfo files is discouraged but not dangerous in itself. However, there is no plausible explanations for why the number of recipients seeking our further information should be higher in the group that already received more detailed explanations, and why they should only apply to letters, not to emails.

A final explanation may be that the messages are read by a different group of people: maybe the people who read and respond to email messages are from a different department than those acting upon letters. This is plausible, as it is likely that more organizations offered a specific *email* address for technical issues than offered a special *postal* address. It may thus be the case that a higher percentage of emails was sent to technical contacts than was the case for letters, and that system operators react different to other people in the organization. As we did not label collected addresses with the type of contact, we cannot determine if the behavior of technical and general-purpose contacts is different, and it also remains unclear why only the PHPInfo group should be affected by this difference. Regardless of the mechanism creating these observed differences, the results highlight that more work is needed to understand the perspectives of system operators that receive unsolicited notifications.

6.5 LIMITATIONS

We now discuss limitations of our study.

A few issues can potentially limit the external validity of our results. First of all, most vulnerabilities in our dataset are PHPInfo leaks, which are arguably the easiest to remediate. However, we've seen that their remediation rates are only marginally higher than that of other vulnerabilities (cf. Figure 10(b)), which limits the impact of this imbalance. The lower percentage of PHPInfo vulnerabilities in the LETTER+ATK group (83.7 vs. 86.2 to 88.7 %) poses another imbalance in the dataset. However, due to the aforementioned low differences in remediation rates between the different vulnerability types, it may contribute to but cannot explain the observed differences between LETTER and LETTER+ATK.

Our dataset had multiple imbalances, the effects of which should be minor.

Our study suffers from observer effects, as we did not attempt to hide that our messages were sent as part of a study. This may have caused recipients to (consciously or unconsciously) alter their behavior. Future studies should consider obfuscating the nature of the notifications to avoid these effects, as long as the deception is revealed after the end of the study. In this case, it is advisable to seek ethical review.

The study is affected by observer effects.

Our dataset is geographically limited to German websites. This increases the availability of contact information on websites due to German legislation requiring its disclosure. It may also promote name recognition of the sending institution, thereby influencing the remedi-

The study was limited to German websites, leading to a higher availability of contact information and increased name recognition.

ation (assuming the identity of the sender has an effect on remediation rates, which is questioned by previous studies [18, 98, 117]).

The coming-into-force of the GDPR at the time of the study may have had an unquantifiable effect on the results.

Finally, we conducted the study three weeks after the GDPR came into effect. If this event had any temporally limited effects on system operators, for example due to an increased concern for data protection, it may skew our results. We are unable to quantify the effects of this.

6.6 CONCLUSION

This chapter described a notification experiment evaluating message medium and -content.

In this chapter, we described our randomized controlled notification experiment with 1359 German website operators affected by unintentional information leaks. We used this dataset[12] to compare the effectiveness of notifications using two different mediums — email and letters — and the inclusion of scenarios describing potential attacks that result from these leaks. We collected contact information manually, which resulted in low bounce rates (less than 2 %).

Notification were effective, especially for letters. Changing the message led to inconclusive results.

The overall remediation rate reached 48.9 % for notified recipients over a period of a month, compared to 4.3 % for an unnotified control group in the same timeframe. The use of letters as an alternative notification channel proved to be successful, in some cases increasing remediation rates by 25 percentage points. On the other hand, descriptions of attack scenarios failed to improve results for the email group, and reduced the overall remediation rates for letters. This trend seems to be driven by a decrease in remediation rate for the most numerous vulnerability, while other vulnerabilities saw increased remediation rates from attack descriptions.

These results raise a number of questions that we attempt to answer with a third notification study, described in the next chapter.

These results leave us with a number of questions. First, given that the only previous study using letters reported only marginal improvements [98], can our results be replicated on a larger dataset? Second, if attack scenarios do not reliably improve remediation, what other arguments could be used? Given the prominence of the GDPR at the time of this study, can legal arguments provide an incentive, as proposed by Çetin *et al.* [18] and our own study concerning health insurance websites? Third, given that several recipients mentioned trusting us because they recognized the sender, does the sender of notification messages really have little to no effect, as claimed by previous studies [18, 98, 117]? Fourth and finally, what are the views of notification recipients? Are unsolicited notifications desireable? How should they be designed and delivered? To answer these questions, we design and conduct another notification study, which we describe in the next chapter.

12 The used messages, code and data underlying the study can be found on Zenodo: https://zenodo.org/record/4817464 (last accessed 2021-07-15).

7

ALTERNATIVE SENDERS AND NON-TECHNICAL ARGUMENTS

In the previous chapters we have shown that utilizing competition only has very limited effects in driving changes to the privacy properties of websites, and even in the area of security, where the interests of users and operators are more aligned, notifications still led to less than half of the notified website operators remediating the reported issue. This raises the question of what other incentive may convince operators to remediate, if the security of their website is not sufficient. Given the near-panic caused by the GDPR coming into effect, which saw companies shut down rather than risk the fees imposed for violations[1], *compliance* may be the answer. This idea that legal liability may be a powerful motivator was previously proposed by Çetin *et al.* [18]. It is also supported by the measurements conducted as part of the first study, which saw significant changes on websites when the GDPR came into force (cf. Section 5.3.4.1). At the same time, the previous study has raised a number of additional questions (cf. Section 6.6), for which we want to collect additional data.

Our previous studies have shown that competition and security arguments have only limited effectiveness.

Legal requirements / liability may serve as an alternative motivator.

To evaluate these, we conducted a third and final study, which will be described in this chapter. We begin this chapter with an overview of the study before continuing with a detailed description of the study design, which includes three components working together to answer our research questions. We then give the detailed results of our study and put them into context, before ending the chapter with a discussion of the limitations of our study and an overall conclusion.

We describe a study that investigates this question in more detail.

7.1 OVERVIEW

Once again, we first discuss the questions our study aimed to answer before going into more detail about the technical and legal background of the misconfiguration under study. We close with a description of the dataset underlying the study.

We discuss the research questions and give necessary background information.

7.1.1 *Research Questions*

With this study, we sought to investigate a number of different questions. First, we aimed to strengthen the evidence for the results of the previous study by repeating the comparison of *email* and *letter* as the contact medium. Second, given that several respondents in the previ-

We investigated message medium, sender, and framing of the problem.

1 See, for example, https://money.cnn.com/2018/05/11/technology/gdpr-tech-companies-losers/index.html, last accessed 2021-03-05.

ous study mentioned trusting the messages because they came from a university (Section 6.4), we sought to replicate or refute previous studies that claimed that the sender of the message does not have a significant impact on remediation rates [18, 98, 117] by comparing three different senders: a *private individual* as well as a *computer science* and a *legal* research group. Third, given the effectiveness of regulatory changes observed in the first study (cf. Section 5.3.4.1), we investigated legal *compliance* as a new framing to incentivize remediation, replacing the security framing of the previous study.

We also collected data using a self-service tool, a survey, and analysis of recipient responses.

Outside of the three core variables of the notification study, we aimed to collect additional data on the use of a self-service check tool that allowed recipients to validate their remediation attempts. We also sought to gain further insight about the perspectives of the recipients by sending them a survey at the end of the study and conducting a more detailed analysis of their responses to our messages.

7.1.2 *The Issue*

We chose a Google Analytics misconfiguration for our study.

To test a compliance-based framing, we needed an issue that has legal implications while being automatically detectable at scale. For this purpose, we used a misconfiguration of the Google Analytics web analytics software, for which we already had a proof of concept detection system implemented in the PrivacyScore platform (cf. Section 3.1.1), which also showed the problem to be wide-spread and thus a promising avenue for such research. We first describe the issue from a technical perspective before discussing the legal aspects that make it suitable for our study.

Google Analytics is a web tracking service.

GOOGLE ANALYTICS BASICS Google Analytics[2] is a web tracking platform operated by Google. It allows website operators to gain insight into how many users visit their website, which sites referred them, and how long they stay on the website, among other features. It can also be integrated with Google's advertising platform[3], increasing its value for websites that rely on advertising to attract customers.

It is embedded and configured using a JavaScript snippet.

Google Analytics is used by including a JavaScript-based library in the website. This JavaScript code is then configured with a tracking ID and any additional settings, before sending *events* to the Google Analytics servers. The most common event is a page view, although other events (like adding items to a shopping cart) can also be tracked.

2 See https://analytics.google.com/, last accessed 2021-03-23.

3 See https://marketingplatform.google.com/about/resources/linking-analytics-and-ads-solution-to-todays-marketing-challenges/, last accessed 2021-03-23.

GOOGLE ANALYTICS IP ANONYMIZATION One of the settings that website operators can activate is the so-called *IP anonymization*[4]. If this setting is activated, the parameter `aip=1` is added to all tracking requests, which instructs the Google servers to "anonymize" the IP address of the visitor by setting the last octet of the IP address to zero (i.e., 130.83.183.199 will be turned into 130.83.183.0). For IPv6 addresses, the last 80 bits are set to zero instead. Enabling IP anonymization requires changes in the JavaScript code that configures the Google Analytics library, and the exact methods differ between different versions of the library. This makes the process error-prone (cf. Appendix C.1).

It offers an "IP anonymization" feature which truncates the IP addresses of visitors before tracking.

Such an anonymization likely has only a small impact on the real-world privacy of website visitors. However, this issue has two distinct advantages that make it suitable for our study: It can be detected automatically (by checking for the `aip=1` parameter), and not activating the feature is considered a data protection law violation in Germany.

The presence or absence of this feature can be reliably detected remotely.

IP ANONYMIZATION AND DATA PROTECTION LAW Even before the GDPR came into effect, using Google Analytics without IP anonymization was considered a violation of data protection law[5]. Under the GDPR, this requirement has been upheld in a 2019 decision of a German sub-court [66].[6]

Activating this feature is mandatory under German data protection law.

Enforcement of the GDPR usually falls to the data protection authorities. However, German competition law also allows competitors of a non-compliant company to send a written warning with costs ("Abmahnung"). As this practice has seen some misuse in the past, German website owners are very conscious of any issue that may result in such a warning.

Competitors of non-compliant websites can attempt to enforce this requirement.

We note that under the GDPR, website owner also bear joint responsibility for any content they embed into their website from third parties [22]. This means that even Google Analytics code used by third party providers (like widgets, CMSs or advertising networks) fall under the responsibility of the website owner. This makes the issue particularly suitable for our purposes, as it ensures that, regardless of the source of the misconfiguration, attributing the legal responsibility

Even if the non-compliant code comes from third-party content embedded in the website, the website operator is legally responsible.

4 See https://support.google.com/analytics/answer/2763052?hl=en, last accessed 2020-12-11.

5 See https://web.archive.org/web/20131205221149/http://www.datenschutz-hamburg.de/news/detail/article/beanstandungsfreier-betrieb-von-google-analytics-ab-sofort-moeglich.html for a German press release on the topic (original source unavailable due to link rot), last accessed 2021-03-05.

6 After the end of the study, the German association of data protection authorities released an opinion that recommends, but no longer requires, IP anonymization for German websites. However, it also requires explicit consent before any data is sent to Google, implying that any website that our scanners find to be using Google Analytics are in violation of the proposed guidelines, regardless of the anonymization status. See https://www.datenschutzkonferenz-online.de/media/dskb/20200526_beschluss_hinweise_zum_einsatz_von_google_analytics.pdf for further details, last accessed 2021-05-17.

for a misconfiguration to the operator of the tested domain will always be correct.

7.1.3 *The Dataset*

We describe the dataset collection method.

For our final notification campaign, we needed to assemble a dataset. We describe the individual steps of the process here.

The dataset contained German websites collected from Wikipedia and internet toplists.

DATA SOURCES We once again relied on data from two sources. For the same reasons outlined in the previous chapter, we focussed our study on websites based in Germany by filtering for the German TLD `.de` (further filtering was performed later in the process). Firstly, we again used data collected from the Wikidata platform, from which we obtained 32 782 domains with the correct TLD. We augmented this dataset with a set of merged and deduplicated internet toplists collected by Scheitle *et al.* [94], filtered by TLD, resulting in 1 265 750 additional domains being added to the dataset, for a grand total of 1 298 532 domains.

DETECTING THE MISCONFIGURATION To detect the misconfiguration, we developed and deployed a *compliance checker* based on the Chrome browser and controlled via the Chrome DevTools protocol, a platform we have prior experience with due to its use in the PrivacyScore scanning stack. The scanner visits websites, hiding its nature as an automated scanning system with a modified user agent, and executes JavaScript and other code like a regular browser, logging all traffic. If traffic to Google Analytics servers is found, it injects custom JavaScript to determine the configuration of the Google Analytics library. The data obtained through this custom JavaScript is not used to determine the compliance status (which can be more reliably inferred from the network traffic). However, it can be used to determine certain types of the misconfiguration, like activating IP anonymization only after the tracking request was sent (cf. Appendix C.1). It also provides some additional information for users of the self-service scanning tool (discussed in Section 7.2.5)

We detected the vulnerability using a custom scanner based on the PrivacyScore stack.

We used the scanner to check the compliance status of all websites in our dataset, finding 3070 (9.36 %) non-compliant websites from the Wikipedia- and 161 984 (12.8 %) from the Toplist dataset. These numbers should be considered a lower bound, as the scanner does not interact with consent forms on the website. Thus, a *cookie consent* form that only embeds a non-compliant Google Analytics after the dialog has been confirmed would not be detected by our system.

We found large numbers of non-compliant websites.

GATHERING CONTACT INFORMATION We selected a subset of these websites for our study by first visiting all non-compliant websites from the Wikidata group and attempting to collect address information

from the imprint. We also categorized them into different groups (company, individual, public sector, ...), which allowed us to avoid biases later in the experiment. Of the 3070 non-compliant Wikidata websites, 1590 were found suitable for the study and included (the criteria for this will be discussed below). To achieve a larger sample, we then repeated the process for a random sample of 5000 non-compliant websites from the toplist dataset. 91 of these were already present in the Wikidata dataset and thus not considered again.

We manually collected contact information and other metadata for a subset of these websites.

The address collection was performed independently by three different researchers for each website. Disagreements were resolved by majority votes or, if no majority existed, discussion inside the team. The process of collecting and validating this data took over 500 person-hours, during which we also performed periodic rescans to remove sites that had become compliant in the meantime.

Three researchers independently collected the data for each website.

During and after the address collection, we excluded a total of 3225 websites from the study for several reasons: The most common issues were that no contact information was found (about 20 % of excluded domains) or the website operator resided outside of Germany (again, about 20 % of excluded sites). Some websites were unreachable at the time of the data collection and excluded (about 10 %). We also excluded websites of politicians to avoid cross-contamination with a different study (less than 1 % of excluded sites). Finally, we scanned all websites again before the messages were sent and excluded all websites that either became compliant or went offline in the months between address collection and the initial notification, which accounts for the remaining excluded websites (approx. 30 %).

We excluded a number of websites from the study for different reasons.

Similarly to the previous study, we attempted to merge websites run by the same operator by deduplicating based on the contact information (email and postal address). We only merged websites that are obviously related, i.e., had identical or very similar addresses (e.g., identical street address and company name only differing by a suffix like "GmbH"). Our final dataset contains 4754 websites run by 4594 owners.

We deduplicated owners of websites based on contact information.

7.2 STUDY DESIGN

We now describe the core experimental setup. We begin with a discussion of the experimental factors that we varied as part of the study, before discussing how recipients were assigned to groups. We then describe the timeline of the experiment, and continue with a discussion of the three core data collection mechanisms used in this study: the monitoring system, self-service tool, and survey of recipients. We close with a discussion of the evaluation methodology and the ethical issues surrounding our study.

We now describe the experimental setup, including timeline, data collection, and evaluation methodology.

7.2.1 *Experimental Factors*

We considered medium, sender and framing as experimental factors.

We considered three experimental factors: The notification medium, the sender, and the framing. We discuss each in more detail below.

Like in the previous study, we compared letter and email.

NOTIFICATION MEDIUM As in the previous study, we were interested in determining the effect of the notification medium. Once again, we compared letters (LETTER) and emails (EMAIL). Emails were again sent in plaintext.

We compared three senders: a private individual, computer science research group, and legal research group.

SENDER We evaluated three different senders for the messages: a private individual with an interest in privacy (CITIZEN), a computer science research group (UNI-CS), and a legal research group (UNI-LAW). Max Maass posed as the private individual. The Secure Mobile Networking Lab at TU Darmstadt served as the computer science sender, and the *Lehrstuhl für Öffentliches Recht, Informationsrecht, Umweltrecht, Verwaltungswissenschaft* headed by Prof. Spiecker at the Goethe-Universität Frankfurt as the legal research sender.

We took measures to ensure that the message looked plausible for the purported sender.

For the two university senders, emails were sent using a purpose-specific email account (*notification@group.university.tld*) hosted on university infrastructure. Letters used regular stamps (instead of automatically printed postage), as printed postage was deemed unrealistic for a private sender. They used the regular postal address of the university and its official letterhead. The letterhead and email signatures contained the full address of the research group and an email and a phone number, including a weekly time window when telephone support would be provided. Max Maass used his home address for letters and created a new email account with a commercial German email provider for the study. His letters did not contain a phone number, and emails contained neither phone number nor postal address, as it was judged unlikely that a private individual would include them in such a message.

We provided recipients with a contact phone number for the two institutional senders.

Finally, we compared three framings: privacy concerns, GDPR violation, and GDPR violation including the potential fines.

FRAMING Finally, we used three different framings: a general privacy concern without referencing any regulation (PRIVACY), a notification about the GDPR violation (GDPR), and an extended version of the latter that also mentions the potential fines for GDPR violations (GDPR+FINE).

The privacy framing served as a baseline to compare the others against.

The PRIVACY message was intended as a baseline, as it contains no special attempts to motivate remediation, aside from a general argument that not enabling IP anonymization is bad for the privacy of the website visitors. The GDPR and GDPR+FINE framings investigated the added effectiveness from providing a legal argument and referencing potential fines, respectively. As no precedent for the expected fines for this violation existed, we chose to cite the maximum sanctions allowed

by the GDPR for the GDPR+FINE framing. The text of the different messages is given in Appendix C.3.

7.2.2 *Group Allocation*

We conducted our study with a *full factorial* design, i.e., all combinations of medium, sender and framing were used, in addition to a control group (CONTROL) which did not receive any notifications. Recipients for whom contact data for only one communication channel could be found in the imprint are randomly assigned to one of the groups that used this channel. This was the case for 87 email-only and 152 letter-only recipients (approximately 6 % of non-CONTROL recipients). The other recipients were assigned to a random group (including the control group). This sampling was stratified by the type of website (company, individual, public sector, etc.) to avoid introducing biases. As many previous studies considered emails, we decided to emphasize letters in our study by assigning twice as many recipients to LETTER groups than to EMAIL groups. This allowed us to collect more data on the behavior and perspectives of letter recipients.

We used a full factorial controlled experimental design. Group assignments were randomized with a few exceptions.

We stratified based on the type of the website to avoid biases.

7.2.3 *Experiment Timeline*

In total, we sent up to three messages to notification recipients.

We sent three sets of messages as part of the study.

INITIAL NOTIFICATION The first message was the initial notification, which was sent to all (non-control) recipients. Email notifications were sent from the 1st to 5th of July 2019, spread over multiple days to avoid triggering rate-based spam filters. To ensure that they would arrive at approximately the same time, letters were sent on the Friday of the previous week to compensate for postal delivery times.

The initial notification was sent in July 2019.

REMINDER Recipients that had not remediated by the 25th of July 2019 received a reminder message on the same medium. Aside from a small number of website operators that had complained about our messages and asked not to be contacted again, we also excluded all recipients for whom the delivery of the initial message failed from the reminders. Recipients that had already contacted us but not become compliant yet received a hand-crafted reminder message if one was judged appropriate based on the previous interactions. Reminder emails were sent on the 1st and 2nd of August 2019, while letters were only sent on the 6th of August due to organizational delays.

We sent reminders to all unremediated sites in August.

Human error in the creation of the reminder messages led to all LETTER – UNI-LAW recipients receiving the GDPR+FINE framing. This mistake may skew the results, and will be discussed in more detail in Section 7.3.1.2.

A mistake in letter creation led to incorrect reminders being sent to some recipients.

The final message was sent in October and invited recipients to participate in a survey.

DISCLOSURE AND SURVEY Finally, on the 1st of October we sent a debriefing message that revealed that the messages were part of a scientific study, and contained a link to a survey with a number of questions for system owners, which will be described in Section 7.2.6. The message also offered recipients the option to opt out of the study. The control group received a separate debriefing message that informed them that their website had been part of a study, to give them the opportunity to remediate the issue as well.

7.2.4 *Monitoring*

We scanned all websites that were part of the study four times per day.

During the data collection timeframe, we used the compliance scanner discussed in Section 7.1.3 to conduct four scans of all study websites per day. This allowed us to determine when the misconfiguration was remediated. Scanning multiple times a day also allowed compensating for scan errors or other factors that can influence individual scans. For example, the multiple daily scans revealed that one website served different content during the day and night, and only one of the two versions was affected by the misconfiguration. Thus, the larger amount of scans increases our confidence in the correctness of our results.

7.2.5 *Self-Service Tool and Support*

We provided a self-service tool and offered limited support via email and phone.

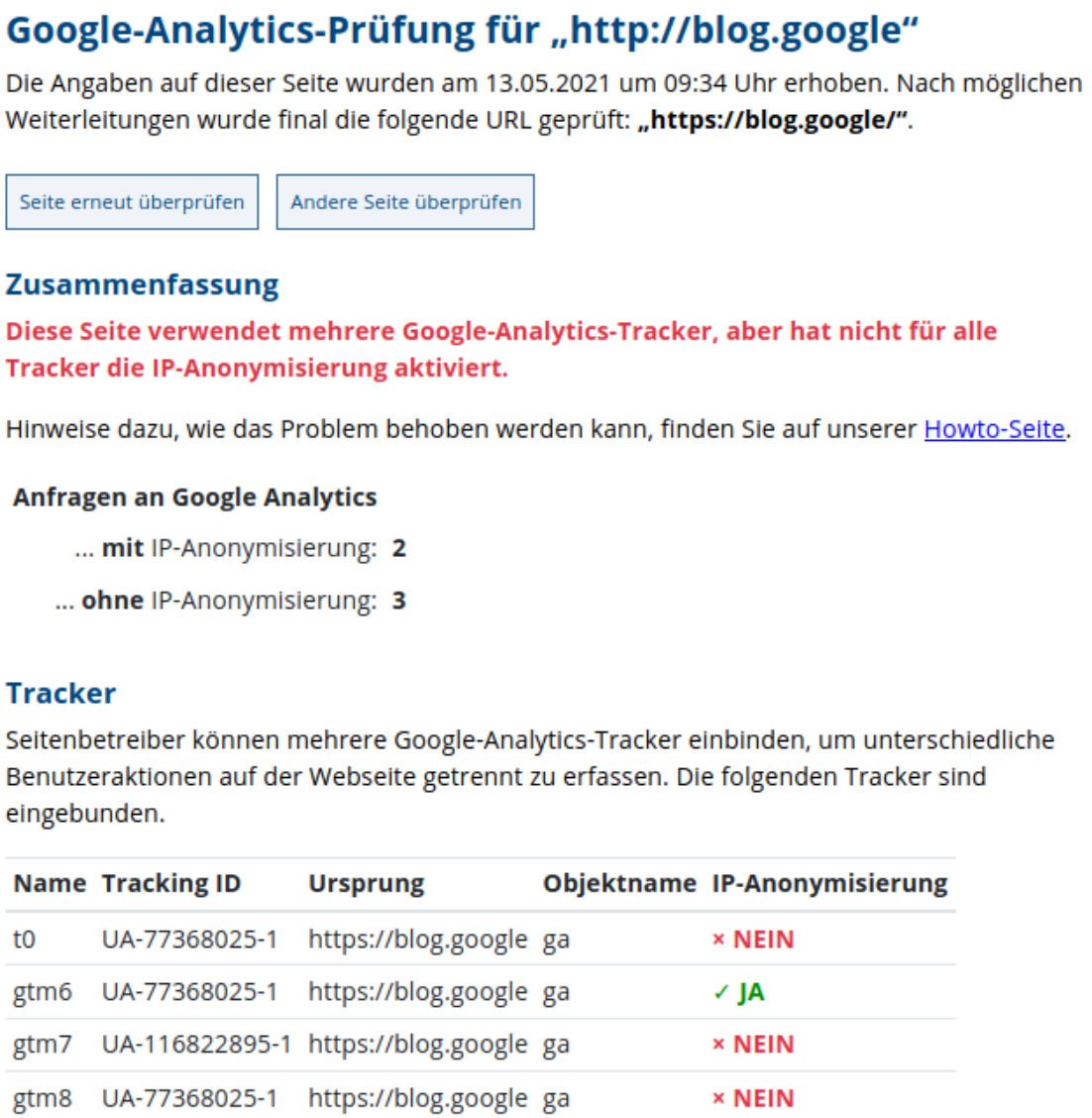
We supported system owners in remediation in two ways: we linked to a self-service scanning tool that allowed them to validate their remediation attempts, and we offered limited support via email and phone.

We operated and linked to a public self-service tool that was not outwardly affiliated with any of the senders.

SELF-SERVICE TOOL Scanning tools are often requested by notification recipients [17, 69, 117], although their effect on remediation rates is reportedly low [17]. For our study, we decided to offer a public scanning tool, called *Check Google Analytics* (CheckGA), that allows recipients (and everyone else) to verify the status of Google Analytics on their website. As we were using multiple different senders, we did not want to associate the tool with one of the involved individuals or institutions to avoid biasing the results. Thus, we hosted the tool under the domain of the research group "*Privatsphäre und Sicherheit in Informationssystemen*" at the Otto-Friedrich-Universität Bamberg, which also operates the PrivacyScore platform and is not outwardly affiliated with any of the senders. To ensure the tool could be found using search engines, we linked it from the homepage of the group and explicitly submitted it for indexing by popular search engines. Over the course of the study, some recipients also shared the tool on social media or wrote blog posts about it, further increasing its reach.

Some recipients shared the tool on social media.

The tool provides users with a detailed report about the presence of Google Analytics on their website (cf. Figure 12). The report includes



Google-Analytics-Prüfung für „http://blog.google“

Die Angaben auf dieser Seite wurden am 13.05.2021 um 09:34 Uhr erhoben. Nach möglichen Weiterleitungen wurde final die folgende URL geprüft: **„https://blog.google/“**.

Seite erneut überprüfen | Andere Seite überprüfen

Zusammenfassung

Diese Seite verwendet mehrere Google-Analytics-Tracker, aber hat nicht für alle Tracker die IP-Anonymisierung aktiviert.

Hinweise dazu, wie das Problem behoben werden kann, finden Sie auf unserer Howto-Seite.

Anfragen an Google Analytics

... **mit** IP-Anonymisierung: **2**

... **ohne** IP-Anonymisierung: **3**

Tracker

Seitenbetreiber können mehrere Google-Analytics-Tracker einbinden, um unterschiedliche Benutzeraktionen auf der Webseite getrennt zu erfassen. Die folgenden Tracker sind eingebunden.

Name	Tracking ID	Ursprung	Objektname	IP-Anonymisierung
t0	UA-77368025-1	https://blog.google	ga	× NEIN
gtm6	UA-77368025-1	https://blog.google	ga	✓ JA
gtm7	UA-116822895-1	https://blog.google	ga	× NEIN
gtm8	UA-77368025-1	https://blog.google	ga	× NEIN

Figure 12: Part of a scan result by the self-service tool (cropped to exclude list of requests).

a list of detected tracker objects with their settings, what their origins are (as some third-party content like Spotify widgets include Google Analytics), and a list of all tracking requests sent to Google Analytics with all parameters. It also contains a help page with more information on how IP Anonymization can be enabled for the different Google Analytics libraries, and a list of common pitfalls and mistakes (cf. Appendix C.1).

It provided helpful information to allow operators to find the source of the misconfiguration.

For the purpose of the study and future evaluations, we instrumented the system with data collection capabilities. We collect the scanned URL, the time of the scan, the truncated IP of the user (with the last octet removed) as well as the TLS session ID [28], which we use to link scans performed by the same user [103]. This allowed us to determine sets of websites that may be operated by the same user

The tool was instrumented for data collection for the purpose of the study.

(as one administrator may scan all their websites to validate their compliance). The tools' privacy policy documents this data collection.

We provided limited one-on-one support.

PERSONAL SUPPORT We also gave limited personal support via phone, email and occasionally letter. This support mostly came in the form of assuring recipients that the message is authentic and resolving complaints and misunderstandings. However, we also provided basic troubleshooting support if recipients contacted us with questions on the configuration. The interactions with the recipients are described in more detail in Section 7.3.3.

7.2.6 *Survey*

Finally, we conducted a survey of recipients, asking their perspective on issues of trust and problem awareness and -solving.

To gain further insight into the perspectives of the recipients, the final message of the study invited them to answer a survey. The survey was hosted on the platform *Soscisurvey*[7] and contained questions about the recipients' perception of our notification messages, their problem awareness and -solving strategies. Additionally, we collected information about their awareness of other tools geared towards system operators, asked if they would like to receive notifications in the future and via which medium, and requested basic information about their affiliation (company size, number of administrators, etc.). The surveys were tailored for the individual groups (based on medium, sender, framing, and compliance status at the end of the study timeframe) and contained 17-21 questions, depending on the group. Results were analyzed using SPSS[8] for basic statistical data and MAXQDA[9] for qualitative analyses of the open replies. A translated version of the survey questions can be found in Appendix C.4.

Different experimental groups received different survey links.

7.2.7 *Evaluation*

Our evaluation consists of data cleaning followed by survival analysis.

Our evaluation process for the data collected by the monitoring system had two basic phases. We began with a data cleaning step, followed by a detailed analysis of the different groups using survival analysis. We describe both phases in more detail here.

We excluded websites that forwarded to many different domains over time.

DATA CLEANING The data collected by the monitoring system and information received from website operators revealed several special cases that warranted excluding websites from the dataset. Firstly, we found that some domains were operated by advertising agencies. These so-called *traffic services* forwarded visitors to the highest bidder and thus frequently changed the domain they were forwarding to. As

7 See https://www.soscisurvey.de/, last accessed 2021-03-23.
8 See https://www.ibm.com/analytics/spss-statistics-software, last accessed 2020-12-18.
9 See https://www.maxqda.com/, last accessed 2020-12-18.

our monitoring system follows redirects, such traffic services resulted in the system scanning different websites at different times, leading to incomplete data for the individual domains. We thus excluded 31 domains that forwarded to three or more different domains over the course of the data collection timeframe.

We also excluded websites hosted on the free tier of Wordpress.com.

Secondly, discussions with system operators revealed that websites hosted on the free tier of the hosting platform *Wordpress.com* contained Google Analytics code controlled by the platform. They were thus outside of the control of the system operator, and IP anonymization had to be enabled by Wordpress.com, leading to correlated remediations.[10] After validating that all websites containing the Wordpress.com tracker (which can be identified by its tracking ID) did not contain any other trackers, we excluded all 22 websites from the study.

Finally, we excluded two mis-labelled domains and four whose operators opted out of the study.

Thirdly, we found that two domains had been incorrectly labelled as German websites and are in fact operated by companies outside of Germany. We excluded these domains from the evaluation. Finally, we also excluded four domains at the request of their operators.

Our evaluation used survival analysis.

SURVIVAL ANALYSIS To evaluate the effectiveness of our notifications, we utilized a method called *survival analysis*, which was used by several prior studies [15, 18, 68, 108, 117]. This method is particularly suitable to the purpose, as it can operate on data where the event of interest (i.e., a problem being remediated) may not have occurred at the time of the analysis. This type of data is called *right-censored data* in the terminology of survival analysis. It relies on estimators like Kaplan-Meier [56] to derive a function $S(t)$ that describes the probability of an event not having occurred at time t (e.g., that a misconfiguration has not been remediated after 20 days). The methodology comes from the medical field, where it is commonly used to compare the effect of different therapies on mortality (hence the name). While in the medical field a high survival rate is desireable, for our case, a lower survival rate is preferable (as it describes websites remediating the misconfiguration). We used the Python library *lifelines*[11] for our evaluation. In addition to the survival rate, this library also provided us with a confidence interval that allowed us to determine the (un)certainty of the estimation.

It attempts to estimate the likelihood of remediation over time.

As one operator may run multiple websites, we needed to account for these correlated fixes in the evaluation.

Our study design put an important twist on the evaluation: We considered multiple websites that are operated by the same person. Thus, their remediations were unlikely to be independent of each other, as an operator that fixes one website is very likely to also fix other websites under their control (see Appendix C.2 for a more detailed discussion and evaluation of this issue). However, we wanted to avoid merging multiple websites run by the same operator into a single

10 Following interactions with one system operator, Wordpress.com activated IP anonymization for all their customers during the study timeframe.

11 See https://lifelines.readthedocs.io/en/latest/, last accessed 2020-12-18. Our evaluation uses version 0.25.4 of the library [26].

entry (as we did in the previous study), as this would lead to a loss of information, and may also result in edge cases where defining the time of remediation becomes difficult. Instead, we used a weighted Kaplan-Meier fit [86] to ensure that every *operator* (instead of every website) has the same impact on the results. We defined the weight w for each individual website as $w = 1/|G|$, where $|G|$ denotes the number of websites run by the same operator. Through this change, we now considered the impact of our notifications on *operators* instead of *websites*, which is arguably the more interesting question to ask.

To address this, we used a weighted fit that ensures each operator has the same influence, regardless of the number of websites they operate.

To reduce the impact of transient scan errors, we considered a website remediated once our monitoring system had observed c consecutive results that showed the website to be remediated (either by enabling anonymization or removing Google Analytics entirely). In counting these consecutive readings, we ignored results that show the website as unreachable unless we obtain c consecutive such readings. This prevented the occasional scan errors caused by a faulty firewall appliance in the network of the monitoring system from impacting the results. For the purpose of our evaluation, we set $c = 5$. We also repeated the evaluation with $c = 3$ and $c = 8$ and found equivalent results, which indicates that the exact threshold does not have a large impact.

Websites were considered remediated after five consecutive scans found the problem to be resolved.

Standard survival analysis can only consider a single remediation event per subject, i.e., once a website becomes compliant, the evaluation assumes that it remains so. We test this assumption in Section 7.3.1.3.

Survival analysis cannot account for websites becoming non-compliant again.

SIGNIFICANCE TESTING Validating if two survival curves are statistically different would usually be done using a log-rank test. However, this test cannot be used if the dataset does not fulfill the *proportional hazard* assumption [13], which our dataset does not.[12] Instead, we used the method described by Klein *et al.* [59] to test the significance of differences between survival curves at a specific point in time (before sending the reminders and at the end of the study timeframe, respectively). This approach uses a $\log(-\log(\cdot))$-transform to improve the statistical power of the test. We corrected for multiple comparisons using the Holm-Bonferroni method [48], which we applied to all statistical tests conducted in this chapter. All reported p-values already include the adjustments made by the Holm-Bonferroni method and can thus be considered significant at $p < 0.05$.

We tested the significance of differences between experimental groups.

All reported significance values were adjusted for multiple tests using Holm-Bonferroni.

12 This assumption is also required to use the Cox regression [23], which would have provided interesting information about the influence of the individual factors on overall remediation.

7.2.8 *Ethical Considerations*

Our study attempted to help operators, but also causes work for them.

As is the case for any study involving human subjects, the ethics of the research must be carefully considered. Our goal is to help system operators become compliant with the law and thus avoid costly fines. However, reading and considering our messages and acting upon them takes time and can also cause stress for the operators. As we believe this to be in their best interest, we consider this acceptable. Contacting system operators "out of the blue" shares some characteristics with spam messages. However, the contact information provided in the imprint is intended for this purpose, and its use should thus be considered acceptable.

It consumed negligible resources on the server.

Our tool could be mis-used, but it was simple enough that others could easily implement it themselves.

On a technical level, our scans consumed server resources equivalent with one normal page load per scan, i.e., four page loads per day. We consider this small resource consumption acceptable, as the costs to the operator should be negligible. Similarly, our scanning tool also does not consume significant resources. As it allows anyone to scan any website, it may be used to identify targets for written warnings with costs (*Abmahnungen*). However, as the underlying technology is very simple, we consider this risk to be acceptable, as anyone planning to misuse the platform would likely be capable of building a similar system themselves without a lot of effort.

Our messages initially deceived recipients, but the deception is lifted at the end of the study.

Our messages did not initially disclose that they were sent as part of a scientific study. This was done to avoid observer effects that may bias the results, but it implies that we are deceiving the system operators. We thus lifted this deception with the final message of the study, which also contained the link to the survey.

The study was approved by two ethics review boards.

Our study was approved by the ethics review boards of the TU Darmstadt and the Otto-Friedrich-Universität Bamberg. The Goethe Universität Frankfurt does not offer a process for ethics approval. However, we sought and received approval from the dean of the law department.

7.3 RESULTS

We now discuss the results of our study.

We now consider the results of our notification campaign. First, we describe the effectiveness of the notifications by discussing remediation / survival rates. Afterwards, we evaluate the usage of the self-service tool in detail. We then briefly discuss the interactions we had with the recipients, before closing with the results of the survey we conducted with website operators.

7.3.1 *Remediation Rates*

To evaluate the effect of our notification, we used survival analysis as described in Section 7.2.7. We recall that this implies that *low* rates are

Table 9: Survival rates in percent for pre- and post-reminder groups and at the end of the study (lower is better). Results marked with † may be impacted by human error, see Section 7.3.1.2. Results are based on 1321 emails and 2644 letters.

Group	Pre-rem.	Post-rem.	End of study
EMAIL	66.3 ± 2.6	75.8 ± 3.1	50.9 ± 2.7
LETTER	55.6 ± 1.9	66.6 ± 2.6 †	39.7 ± 1.9 †
CITIZEN	59.9 ± 2.7	69.0 ± 3.4	43.9 ± 2.7
UNI-CS	61.4 ± 2.7	70.8 ± 3.4	46.0 ± 2.7
UNI-LAW	55.0 ± 2.8	69.5 ± 3.8 †	40.3 ± 2.7 †
PRIVACY	69.6 ± 2.6	75.1 ± 3.2 †	54.7 ± 2.7 †
GDPR	56.6 ± 2.8	69.0 ± 3.7 †	41.9 ± 2.7 †
GDPR+FINE	50.1 ± 2.8	63.3 ± 3.9	33.7 ± 2.6
All notified	58.8 ± 1.6	70.3 ± 2.0 †	43.4 ± 1.6 †
CONTROL	93.0 ± 2.4	97.6 ± 1.7	90.8 ± 2.6

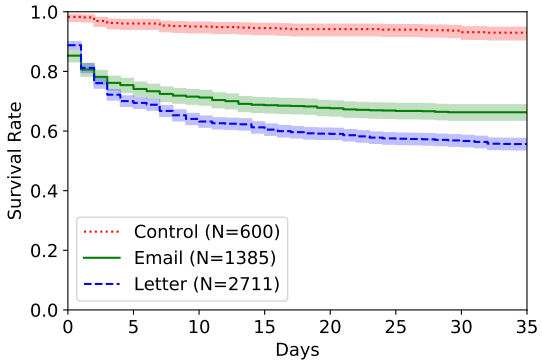
Table 10: Significance levels for comparison of survival rates for groups at different points in time.

	Group	UNI-CS	CITIZEN	UNI-LAW
Pre	CITIZEN	1.0		
	UNI-LAW	*	0.088	
	CONTROL	****	****	****
Post	CITIZEN	1.0		
	UNI-LAW	1.0	1.0	
	CONTROL	****	****	****
Full	CITIZEN	0.920		
	UNI-LAW	*	0.588	
	CONTROL	****	****	****

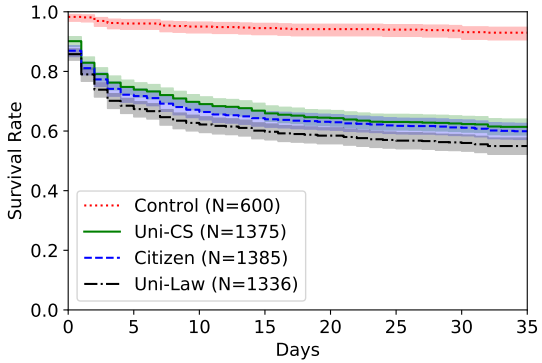
	Group	PRIVACY	GDPR	GDPR+FINE
Pre	GDPR	****		
	GDPR+FINE	****	*	
	CONTROL	****	****	****
Post	GDPR	*		
	GDPR+FINE	****	0.123	
	CONTROL	****	****	****
Full	GDPR	****		
	GDPR+FINE	****	**	
	CONTROL	****	****	****

	Group	EMAIL	LETTER
Pre	LETTER	****	
	CONTROL	****	****
Post	LETTER	****	
	CONTROL	****	****
Full	LETTER	****	
	CONTROL	****	****

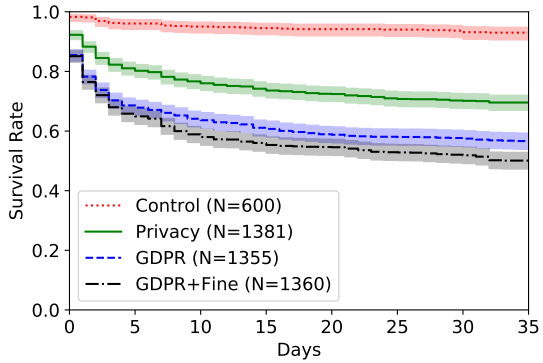
*: <0.05 **: <0.01
: <0.001 *: <0.0001



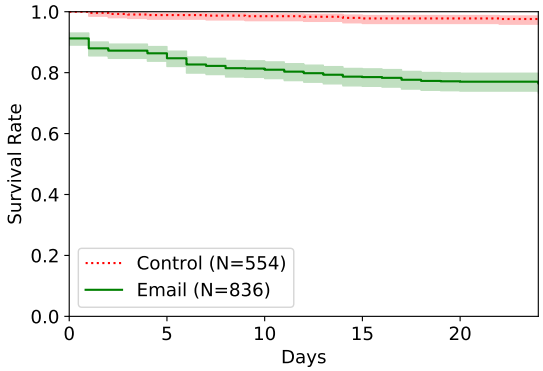
(a) Contact Medium, pre-reminder



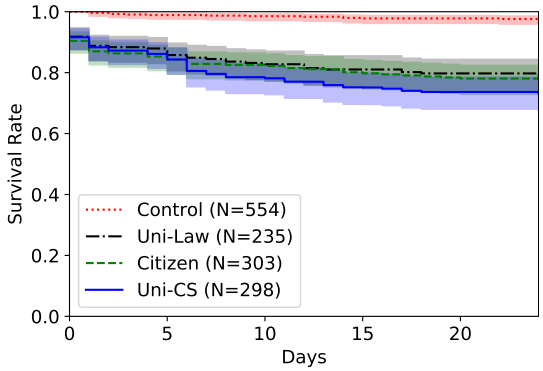
(b) Sender, pre-reminder



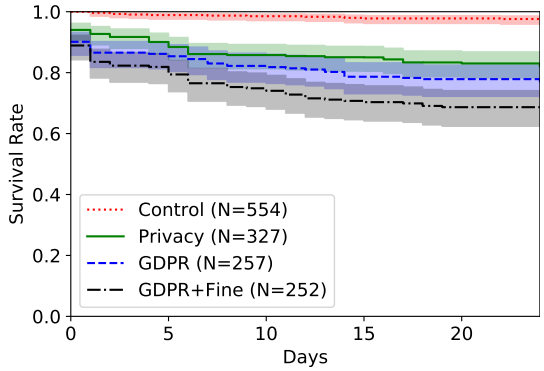
(c) Framing, pre-reminder



(d) Contact Medium, post-reminder



(e) Sender, post-reminder



(f) Framing, post-reminder

Figure 13: Survival rates after initial notification and reminder. Reminder shows data only for Email groups.

Low survival rates are desireable, as they correspond to high remediation rates.

desireable, as they describe a lower probability of a misconfiguration remaining unremediated. The survival rates for both the initial message and the reminder are shown in Figure 13 and Table 9. We discuss the initial messages and the reminders separately before closing with a brief discussion of the used method of remediation (repair vs. removal of Google Analytics) and long-term effects of our messages.

7.3.1.1 *Initial Notification*

Survival times were calculated from the day the message was sent.

As described in Section 7.2.3, the messages were sent at the beginning of July 2019. The survival rates cover a timeframe starting on the day the recipient is expected to have received the message (i.e., the 1st of July for letters, and the day the message was sent for emails), and ending on the day we began sending the reminder messages (approximately five weeks).

The first message was undeliverable for between 3.5 and 5.8 % of recipients.

MESSAGE DELIVERY In the first batch of messages, 48 out of 1337 emails (3.5 %) and 153 out of 2660 (5.8 %) were returned as undeliverable. As always, the number should be considered a lower bound for emails, as not all email servers notify the sender about undeliverable or spam-filtered messages.

Letters were more effective than emails.

MEDIUM As in the previous study, we once again observed letters to be more effective than emails: The misconfiguration survival rate of the EMAIL group after the initial message was 66.3 ± 2.6 %, while the LETTER group saw a survival rate of 55.6 ± 1.9 %, a highly significant improvement ($p < 0.0001$, cf. Table 10 for all significance levels). Like all other treatment groups, both groups significantly ($p < 0.0001$) outperformed the CONTROL group, which had a survival rate of 93.0 ± 2.4 %. Most remediations took place within the first 7-10 days, with only comparatively few in the following weeks until the end of the timeframe.

The law sender was the most effective.

SENDER The differences between the senders is smaller: notifications sent by the CITIZEN sender achieved a survival of 59.9 ± 2.7 %, outperforming UNI-CS (61.4 ± 2.7 %) but being outperformed by UNI-LAW with its survival of 55 ± 2.8 %. Of these differences, only the difference between the two extremes, UNI-CS and UNI-LAW, was statistically significant at $p < 0.05$.

Framing the issue as a legal problem significantly increased remediation rates.

FRAMING Finally, the framings show a very pronounced difference in survival rates: The PRIVACY framing saw a survival rate of 69.6 ± 2.6 %, while the GDPR framing resulted in a much lower survival of 56.6 ± 2.8 %, a highly significant improvement ($p < 0.001$). This is further improved by the GDPR+FINE framing, which had the lowest survival rates of all, 50.1 ± 2.8 %, a statistically significant improvement compared to the GDPR framing ($p < 0.05$).

Medium	Sender	Framing	Owners	Sites	Pre-rem. [%]	Post-rem. [%]	End of study [%]
EMAIL	CITIZEN	PRIVACY	146	163	79.8 ± 7.5	80.7 ± 8.7	63.5 ± 8.4
		GDPR	149	153	63.8 ± 8.2	77.8 ± 10.1	49.0 ± 8.2
		GDPR+FINE	148	159	64.3 ± 8.3	74.1 ± 10.3	48.8 ± 8.3
	UNI-CS	PRIVACY	146	166	82.0 ± 7.5	78.7 ± 9.0	66.1 ± 8.3
		GDPR	149	152	62.4 ± 8.3	74.7 ± 10.8	47.0 ± 8.2
		GDPR+FINE	145	147	61.0 ± 8.5	63.4 ± 11.4	39.3 ± 7.9
	UNI-LAW	PRIVACY	147	149	65.6 ± 8.3	88.1 ± 9.1	55.6 ± 8.4
		GDPR	144	147	65.6 ± 8.3	78.8 ± 10.7	53.1 ± 8.5
		GDPR+FINE	147	149	52.3 ± 8.3	67.6 ± 12.5	35.4 ± 7.7
LETTER	CITIZEN	PRIVACY	294	308	69.2 ± 5.6	70.6 ± 7.1	52.9 ± 5.9
		GDPR	294	304	50.5 ± 5.8	60.9 ± 8.8	33.0 ± 5.4
		GDPR+FINE	292	298	48.4 ± 5.8	59.0 ± 8.9	30.9 ± 5.4
	UNI-CS	PRIVACY	294	302	68.5 ± 5.6	76.7 ± 6.7	55.8 ± 5.9
		GDPR	292	305	54.6 ± 5.9	65.0 ± 8.7	39.8 ± 5.6
		GDPR+FINE	293	303	51.9 ± 5.8	64.4 ± 8.5	35.4 ± 5.5
	UNI-LAW	PRIVACY	293	293	62.5 ± 5.8	70.4 ± 7.5†	44.7 ± 5.8†
		GDPR	288	294	55.6 ± 5.9	68.5 ± 8.2†	41.3 ± 5.7†
		GDPR+FINE	293	304	39.4 ± 5.6	54.7 ± 10.0	23.7 ± 5.0
All notified			3954	4096	58.8 ± 1.6	70.3 ± 2.0†	43.4 ± 1.6†
CONTROL			585	600	93.0 ± 2.4	97.6 ± 1.7	90.8 ± 2.6

Table 11: Survival rates for all groups. Results marked with † are impacted by human error, see Section 7.3.1.2. Remediation rate = 100 - survival.

The contrast between the best and worst groups was very large.

INDIVIDUAL GROUPS To illustrate how these individual differences between the experimental conditions can add to one another, we briefly highlight the best and worst experimental group. The worst-performing notified group (EMAIL – UNI-CS – PRIVACY) achieved a survival of 82 ± 7.5 %, while the best group (LETTER – UNI-LAW – GDPR+FINE) reduced the survival rate to 39.4 ± 5.6 % (cf. Table 11).

Even only considering mails, the differences were still large.

Even if we only consider EMAIL groups due to the cost of letters, the differences are still large: the best email group (EMAIL – UNI-LAW – GDPR+FINE) achieved a performance of 52.3 ± 8.3 %, which is still a 30 percentage-point improvement over the worst group. This shows that the differences between the experimental factors can compound to produce very large differences in remediation rates, which makes it worthwhile to consider them even in cases where sending letters is not an option.

Some websites went offline instead of being remediated.

OFFLINE WEBSITES So far, we have only considered remediation or non-remediation as outcomes. However, some operators chose to take their websites offline instead of remediating.[13] We observed 59 non-CONTROL websites going offline within the considered five-week period, accounting for 1.4 % of these websites. For comparison, six websites (1 %) of the CONTROL group went offline in the same timeframe.

7.3.1.2 *Reminder Message*

We only sent reminders to unremediated recipients where the delivery of the initial message did not fail.

Only website operators that had received the initial message (i.e., it did not bounce) and had not been compliant on the 25th of July 2019 received a reminder message and were considered in the following evaluation (for the control group, we considered all sites that were non-compliant on the 2nd of August). Recipients that had previously responded to us received a hand-crafted reminder, if one was deemed appropriate. Reminder emails were sent on the 1st and 2nd of August, while organizational reasons delayed letters to the 6th of August.

Some reminders could not be delivered.

DELIVERABILITY Interestingly, even though we only contacted recipients where the initial message was not returned to the sender as undeliverable, five out of 809 emails (0.6 %) and 27 out of 1351 letters (2 %) could not be delivered.

Human error prevented us from conducting a detailed analysis of the reminders.

SURVIVAL ANALYSIS As previously mentioned, human error led to us sending the GDPR+FINE framing to all three LETTER – UNI-LAW groups, which contaminated their results. As this contamination impacts many combinations of groups, we only include email messages

13 These cases are counted as non-remediated websites for the purpose of the survival rates, as we cannot be sure that the websites are taken offline as a form of remediation instead of other reasons.

Table 12: Survival S in percent and sample size N of UNI-LAW – LETTER groups after initial notification (i) and reminder (r), survival differences to GDPR+FINE in gray. Results marked with † erroneously received the GDPR+FINE framing.

	Group	S_i	N_i	S_r	N_r
LETTER	GDPR+FINE	39.4	304	54.7	117
	GDPR	55.6 +16.2	294	68.5 +13.8 †	148
	PRIVACY	62.5 +23.1	293	70.4 +15.7 †	169

in the plots in Figure 13, and mark all potentially contaminated results in Table 12 with a † symbol.

We refrained from a detailed analysis and comparison of the different experimental factors. However, we note that the most effective group was LETTER – UNI-LAW – GDPR+FINE with a survival of 54.7 ± 10 % at the end of the study timeframe (after 24 days). Interestingly, the worst-performing group was also a UNI-LAW group (EMAIL – UNI-LAW – PRIVACY), which achieved a survival rate of 88.1 ± 9.1 %. This disappointing result was still an improvement over the control group, which showed a survival rate of 97.6 ± 1.7 %.

The reminder generally had a large impact.

ACCIDENTAL EXPERIMENT: INCREASING THE PRESSURE While the erroneous reminders precluded a more detailed analysis of the reminder messages, they offered us a window into a different question: what happens if the reminder message contains a different argument with (presumably) higher perceived pressure to act? We thus briefly consider the results in this light.

The incorrectly sent reminders allowed us to evaluate if sending a reminder with a legal argument can increase remediation compared to a privacy argument.

Let us first consider an intuitive argument for why such an arrangement may make sense. Receiving a mild first message may lead some recipients to discount the danger of the misconfiguration and refrain from acting upon it. These recipients may have their minds changed by a reminder that spells out the danger in more detail. Leading with the more severe message may scare recipients who would have already acted upon a less severe message (in our study, several recipients complained that the GDPR+FINE message had scared them with the high potential fees), which may lead to undesired over-compliance (people shutting down their website instead of repairing the misconfiguration) or pushback from recipients.

Intuitively, we might have expected this to be the case.

If this argument was accurate, we would expect the remediation rates for the GDPR and PRIVACY groups to be at least equal to that of the GDPR+FINE group. It may even exceed it, as the GDPR+FINE group may have already reached most system operators that are willing and able to remediate under *any* circumstances (e.g., because the remaining operators didn't receive the message or distrust it), while the other two

If it was the case, groups that did not receive the GDPR+FINE *framing should have had a high remediation rate.*

groups may have a larger fraction of operators that would be generally open to remediating, but weren't convinced by the first message.

However, this was not the case.

Looking at the data, this does not seem to be the case: Even though all three groups received identical messages, the remediation rates still followed the trends from the first message: the GDPR+FINE group outperformed GDPR by 13.8 percentage points and PRIVACY by 15.7. They thus mostly maintained the trends from the first message, although the difference between GDPR and PRIVACY was less pronounced for the reminder.

We did not have a control group for this experiment, but it seems to indicate that changing the text of the reminder is ineffective.

As this experiment was unplanned, we did not have a control group to compare the behavior against, rendering any results anecdotal. However, the results indicate that significant changes to the message for the reminder are ineffective. A possible mechanism is suggested by responses we received to our third set of messages, which invited recipients to participate in the survey. These messages led to several reactions from recipients that asked us why we kept sending them messages even though they had already resolved the issue. This indicates that recipients may have learned to recognize our messages by the sender and general layout and stopped reading them in detail, which would make changes to the content ineffective. A possible mechanism that could explain such an effect would be the concept of *habituation* [88] from the field of psychology.

This may be explained by habituation.

Some additional websites went offline during the reminder timeframe.

OFFLINE WEBSITES After the reminder, 31 additional websites were offline, including two from the control group.

7.3.1.3 *Other Aspects of Remediation*

We investigated two other details of the remediations.

We briefly consider two more aspects of the notifications: The type of remediation, and the long-term effectiveness.

So far, we did not distinguish between anonymizing and completely removing Google Analytics, although it can make a large difference in practice.

REPAIR VS. REMOVAL In our evaluation, repairing the misconfiguration and completely removing Google Analytics have been treated as equivalent so far. However, in practice, this is an important difference, for both system operators and website visitors, as removing Google Analytics prevents the website operator from collecting *any* data, while anonymization only slightly reduces the amount. Thus, intuitively, we would assume system operators to prefer remediation over removal, as they presumably have an interest in collecting data about their visitors (otherwise they would not have added Google Analytics to their website in the first place).

Over a third of remediating website operators did so by removing Google Analytics instead of activating anonymization.

Surprisingly, we found that 36 % of website operators that are compliant at the end of the study timeframe completely removed Google Analytics instead of remediating the misconfiguration. As this observation is consistent across the different experimental groups, this does not seem to be related to specific factors of the notification (like the used framing). We manually visited 50 of these websites to validate

Table 13: Survival rate S and CheckGA usage of all (U_a), remediated (U_r), and unremediated (U_u) owners after initial notification and at the end of the study.

Group	S	U_a	U_r	U_u
Pre-reminder	58.8	33.9	65.1	12.5
End of study	43.4	46.9	67.6	19.8
CONTROL (end of study)	90.8	3.1	14.8	1.9

that this result is correct (and not an artifact of the introduction of Cookie Consent banners or similar techniques that may hide the presence of a tracker until the user interacted with the website) and find no false negatives. We discuss this counterintuitive result later in this chapter.

LONG-TERM EFFECTIVENESS A second issue we have so far not taken into account is if the remediation persists over time, as survival analysis cannot handle cases with more than one death event (i.e., remediation) per subject. We thus crawled all websites again in April 2020 (7 months after the end of the study) and checked their compliance status. We found that 78 out of 2224 previously compliant (3.5 %, 6 of 78 from the control group) had become non-compliant again in the intervening time. 38 further sites had become unreachable. We can thus conservatively estimate the long-term effectiveness of our notifications at approximately 95 %.

Remdiating sites mostly stayed remediated over time, only a small percentage became non-compliant again.

In the opposite direction, we saw that of the 2371 non-compliant sites (550 from control group), 438 non-control (24.1 %) and 82 control-sites (14.9 %) were compliant 7 months after the end of the study, with another 63 becoming unreachable. We note that the number of compliant websites may be overestimated due to the rising prevalence of cookie consent systems that block the inclusion of Google Analytics until an interactive banner has been confirmed. If we assume no false detections, this would put the base rate of unnotified remediations at 14.9 % over 7 months, with an increase in the rate to 24.1 % for notified sites in the same timeframe.

About a quarter of sites that were non-compliant at the end of the study became compliant within 7 months after the end.

7.3.2 *Use of Self-Service Tool*

We now discuss the effectiveness of our self-service tool, Check Google Analytics (CheckGA).[14] The tool allows site owners to verify if IP anonymization is activated correctly on their website. Over the timeframe of the study, CheckGA performed 38 485 scans of 14 023 websites.

We evaluate our self-service check tool.

14 At the time of writing, the tool is available online at https://checkgoogleanalytics.psi.uni-bamberg.de, last accessed 2021-01-04.

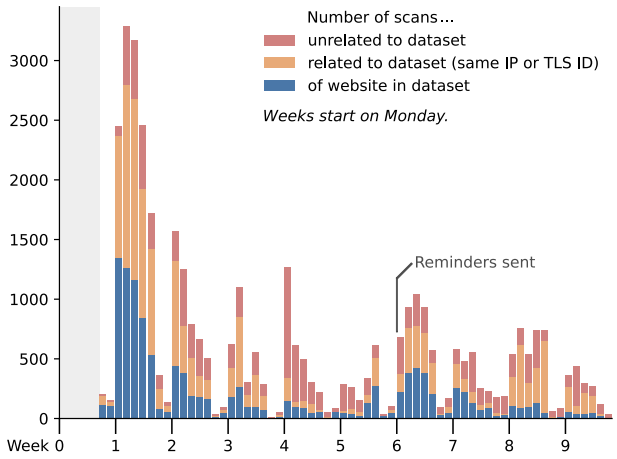


Figure 14: User-initiated CheckGA scans per day.

It was widely used and also scanned many websites that weren't part of the study.

Of these websites, 12 047 were not contained in the study dataset, meaning that they are either other websites operated by recipients of our messages, or websites completely unrelated to our study. We did not widely advertise the tool, however, over the course of the study, multiple people tweeted or blogged about it, recommending it to their colleagues or networks. However, given that only 3.1 % of the CONTROL group was scanned with the tool, we assume that most scans of websites that were part of the study were still performed by their owners and in connection with our study. This would mean that approximately 46.9 % of notified operators used the tool at least once on their site(s). We provide the fractions of owners that used the tool (U_a) as well as the fraction of owners of remediated (U_r) and unremediated (U_u) sites that used CheckGA in Table 13.

46.9 % of the notified operators used the tool at least once.

TIME DISTRIBUTION We give the number of scans per day during the study timeframe in Figure 14, starting on the friday of week 0 when the first letters were sent. We considered a scan to be *part of* the dataset if it either directly scanned a domain in the dataset, or forwarded to such a domain.[15] We consider a scan *related to* the dataset if it did not target a site in the dataset, but was performed by the same user as a scan that was *part of* the dataset, where users are identified by truncated IP and TLS session ID, as described in Section 7.2.5. Any scan not matching either of these two rules was considered unrelated.

We evaluate the distribution of scans over time.

Most scans happened on weekdays, and many users also scanned other websites in addition to their own.

The figure shows that most scans happened on weekdays, with only very few on the weekend. It also shows a large number of related scans, meaning that the recipients either chose to scan other websites under their control, or other, unrelated websites that they are curious about. However, there are also surprisingly many scans that don't

15 For clarity: we considered a domain part of the dataset if it was either part of the original list of domains, or if at least one forward to it from a domain that was part of the study was observed over the course of the study.

appear to be related to the dataset at all, and are thus likely due to word of mouth recommendations about the tool.

SCANS UNTIL COMPLIANCE However, simply scanning the website is not enough if the problems aren't also remediated. As a proxy for the helpfulness of the tool for remediation, we counted the number of scans until *all* subsequent CheckGA scans find the site to be compliant. In the median, the users performed two scans before a website was either remediated or no further scans were performed. The mean was similar for both groups (4.5 vs. 4.16), which indicates that users either managed to remediate quickly, or gave up.

We evaluate if scans also lead to remediation.

In the median, it took 2.22 hours from the first scan to completed remediation. However, the mean was significantly larger at 5.05 *days*. The fastest quartile of sites was compliant within 3.3 minutes, the third quartile over 28 hours. This indicates that while many site owners managed to remediate quickly, many others also needed long amounts of time or took extended breaks from their attempts, a dynamic also observed by Li *et al.* [69]. The latter may also be due to the initial recipient confirming the issue with our tool before passing it on to their web design agency, which took a day or two to begin work and finally used the tool to verify the remediation.

Many websites were remediated quickly after the first scan, but some took a long time.

7.3.3 *Communication with Recipients*

Our notifications led to many conversations with recipients over different media and for a variety of reasons. We briefly summarize the number and types of responses. We received 946 emails (excluding autoreplies), 41 letters, 56 phone calls (a significant fraction of which did not respect the time window provided in the notification message), and two Twitter messages. These messages were sent by 764 distinct recipients, and we responded with 374 emails, one letter, and 12 phone calls. We summarize the most common themes below (the list is not exhaustive).

We analyze the received responses from recipients.

REQUESTS FOR CONFIRMATION 32 recipients (4.2 % of those that contacted us) sought confirmation that the message was authentic. They often (but not always) used a different contact channel than the one they were contacted with, although a small minority simply responded to the suspicious email and asked for confirmation of its authenticity. However, most chose the more prudent path of looking for alternative contact details on university websites and using the email addresses or phone numbers found there. Some even contacted the university department or press contact instead of the involved research group, and were forwarded from there. Many of these requests were friendly and curious, although some were initially hostile, alleging

Some wanted to ensure the message was authentic.

bad intentions (i.e., scams / unsolicited advertising) or complaining that the message was impossible to understand for laypeople.

REQUESTS FOR HELP Another common reason to contact us was to request aid, either in remediating the issue or in verifying that the remediation was successful (a request that was made surprisingly often, considering that we prominently linked to a tool for this exact purpose). In total, we received such messages from 204 recipients (26.7 %). One operator from the CITIZEN group asked us to remediate the issue for them, offering us login information for their web server so we could remediate the issue "if it is that important to you." Another operator asked for support with an unrelated PHP issue on their homepage. We provided remediation instructions and validations upon request, but did not directly assist in remediation.

Others requested help in resolving the issue.

COMPLAINTS Similar to the previous two studies, not all recipients had a positive response to our notification message. In total, 19 (2.5 %) recipients chose to complain in a variety of forms. Some were simply unhappy about the tone of the message or found the perceived threats in the GDPR+FINE framing stressful. Others threatened legal action (in one case contacting the chancellor of TU Darmstadt directly), and one tried to bill us for the time they spent acting on our message. We discussed these cases inside the team and responded with the help of the expertise of our legal collaborators, which proved invaluable. To date, no legal action has been filed against any involved individual or institution.[16] Where possible, we placated these recipients and excluded them from further messages.

Still others complained about the tone or content of the message.

THANKS The largest class of responses is that of simple expressions of gratitude. 260 recipients (34 %) contacted us to communicate their thanks for our notifications. They included simple messages of thanks, but also offers of payments, discounts, or gifts. Some recipients sent unsolicited packages of gifts, including mugs and magazines, but also a donation to one involved university. For ethical and legal reasons, we turned down gifts whenever possible.

Most simply wanted to thank us for informing them.

7.3.4 *Survey Responses*

We describe the results of our survey.

Our final major piece of data collected for this study is a survey of notification recipients, who were invited to participate with the final message of the study, which also contained the information that the

16 One respondent was a *Reichsbürger* who does not recognize the legal basis of the German state or the European Union and pursued legal action through their own court system, alleging our messages to be an "act of aggression with the intent to precipitate an armed conflict". We did not respond to their message.

previous message(s) are part of a study. The methodology is described in Section 7.2.6, the survey questions are given in Appendix C.4.

We received 561 responses, of which we excluded 84 because they did not agree to the consent dialogue at the start of the study (19) or answered less than 50 % of the questions (65). Of the remaining N = 477 responses, 226 completed the study. The N varies for different questions in the evaluation below as the survey does not contain mandatory questions, and some questions are only shown to some groups of recipients. The questions can be grouped into three major areas: problem awareness, trust in the notifications, and problem solving.

We received 561 responses, of which 226 answered all questions.

7.3.4.1 *Awareness*

Surprisingly, only 371 out of 461 (80.5 %) of respondents were aware that Google Analytics was active on their website before receiving our notification. Conversely, 272 out of 461 (58.9 %) had previously heard about the IP anonymization feature, and 58 out of 458 (12.7 %) knew that it was not enabled on their website.

A surprising number of respondents were unaware of Google Analytics or IP Anonymization.

We included a question specifically for respondents who had not yet remediated the issue, asking why it had not been done yet (N = 54, multiple responses possible). The cited problems included not knowing about the problem (22) or lack of knowledge about how to solve it (20), which is surprising, considering that they should have received our previous notifications. It may be that the survey was answered by a different person from the one that received the initial notifications, or that notifications were lost. Others also cited a low priority of the problem (12), lack of time (10), or deeming the notification not serious (6).

Others gave different reasons for not having remediated the issue yet.

7.3.4.2 *Trust*

On a four-point Likert scale, 316 of 460 (68.7 %) respondents (rather) agreed with the statement that the notification made a trustworthy impression. When broken down by experimental factors, the notifications from the UNI-LAW group was perceived as most and from CITIZEN least trustworthy. The differences for the remaining factors were less pronounced (cf. Figure 15).

We asked respondents to rate the trustworthiness of the message.

To collect additional data on the reasons for trust and distrust, we included two open questions which were answered by 377 and 252 respondents, respectively (multiple answers possible). We group the resulting factors into three aspects: formal, content-related, and verifiability.

We also gave them the opportunity to name factors that increased or decreased their trust.

FORMAL FACTORS By far the most cited factor that promoted trust was the sender of the message, which was named by 348 of the 377 respondents. 174 (46.1 % of all responses) explicitly referenced

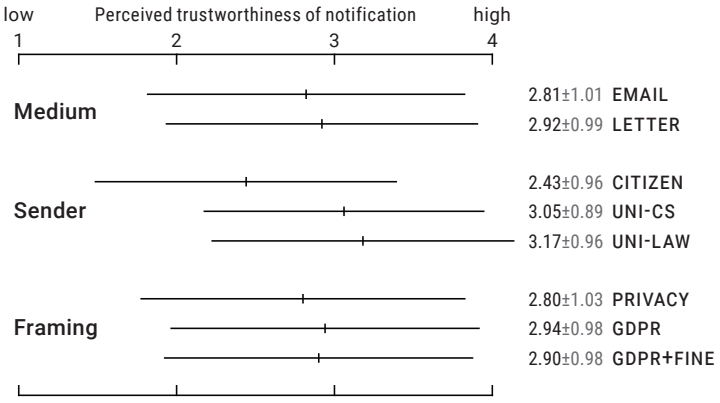


Figure 15: Agreement of website owners with the statement that the notification made a trustworthy impression.

The sender of the message was frequently named as a trust-promoting factor.

The medium and even the design of the message were also cited as trust-promoting by some, but not consistently so.

the fact that the message came from a university, and 44 (11.7 %) mentioned the possibility of contacting the sender as trust-promoting. The correctness of the grammar and spelling of the messages were mentioned surprisingly often (16.1 %; 63 of 377).

25 of the 259 respondents (9.6 %) that received a letter named the medium as trust-promoting. Even seemingly small aspects of the letter like the corporate design with logo and letterhead (18.9 %; 49 of 259) or signature (4.6 %; 12 of 259) increased trustworthiness for some respondents. However, such small factors also decreased the perceived trustworthiness for some — perceived bad wordings (16.3 %; 41 of 252) or layout (18.2 %; 46 of 252) were mentioned as trust-decreasing factors as well, and 12 recipients (4.8 %) stated that receiving a letter was *decreasing* their trust, with one participant wondering "why would anyone even bother to send a letter?"

The correctness of the information and lack of profit motive increased perceived trustworthiness.

CONTENT-RELATED FACTORS Of course, aside from the formal factors, the actual content of the message also had a large impact on its perceived trustworthiness. The most commonly cited factors were the factual correctness and detailed explanations (24.3 %; 94 of 377), as well as the CheckGA tool (14.8 %; 56 of 377). Another common factor was the lack of a profit motive and/or threats in the message (20.2 %, 76 of 377). Conversely, other respondents found the motivation to be unclear (15.1 %; 38 of 252) or thought that the message was a threat or advertising (25.4 %; 64 of 252), decreasing its trustworthiness.

Finally, the verifiability of the information increased trust.

VERIFIABILITY FACTORS The final major factor was the verifiability of the given information. While some recipients claimed not to trust any information from unknown senders (4.4 %; 11 of 252), a larger number rated the verifiability of the information as trust-promoting (31.6 %; 119 of 377). This included both the possibility of verifying the authenticity of the sender (through phone calls, emails, or other chan-

nels), and the correctness of the facts given in the message (through their own research or with the help of experts of acquaintances).

7.3.4.3 *Problem Solving*

Of course, a notification should not only seem trustworthy, but also prove helpful in remediation. We thus included questions about the helpfulness of the notification in problem solving, the usefulness of CheckGA, and the general desires of the recipients when it came to future notifications.

We also evaluated the helpfulness of the messages.

PROBLEM SOLVING Most respondents (77.6 %; 339 of 437) were able to understand the described problem from the notification. More than a third of recipients was able to resolve the issue without outside help (37.8 %), while others asked external service providers (30.9 %) or colleagues (13 %). 10.8 % solved the problem themselves after getting external help (other: 7.5 %, N = 362).

Most respondents could understand the problem based on the notification, and could resolve it.

SELF-SERVICE TOOL The vast majority of respondents (87.2 %; 266 of 305) rated the CheckGA-tool as helpful or very helpful on a four-point Likert scale. 86 respondents stated that they did not use the tool, while 51 did not know about the tool (despite the fact that it was linked prominently in the notification messages). Thus, providing a tool seems to have both increased the perceived trustworthiness of the notifications and helped many operators to remediate.

The self-service tool was generally viewed favorably.

FUTURE NOTIFICATIONS Most operators seem to be happy about having received a notification, with most respondents (88.4 %; 396 of 448) reporting a desire for more notifications in the future. The most-requested contact channel was email (84.8 %), with only 28.2 % requesting a letter. Other channels such as blog posts (3.7 %) or calls (3.2 %) were not considered desireable by many (1.7 % had other preferences like a service portal that required prior sign-up for notifications. N = 401, multiple answers possible). Interestingly, 30.5 % (117 of 305) stated that they would be willing to pay for such notifications, although we note that economic research has found that the self-reported and actual willingness to pay frequently diverge [91, 102, 116]. Thus, a future study could try to validate this experimentally.

Respondents were open to receiving future notifications.

About a third would have been willing to pay for such notifications.

7.4 DISCUSSION

After describing the outcome of the notification experiment and the survey, we now put these results into context. We begin with the observed behavior (i.e., the ground truth effect of our notifications), which allows us to put the self-reported results of the survey into perspective. Finally, we compare and contrast our results with those of prior studies in this field.

We now contextualize the results and draw conclusions.

7.4.1 *Observed Behavior*

Notifications increased remediation compared to the control group, but the performance of the groups varied widely.

Our experiments show that notifications had a significant impact on remediation, with 56.6 % of contacted operators remediating within two months, compared to only 9.2 % of the control group. Remediation frequently took place within the first seven to ten days after message receipt, regardless of other factors of the message. However, there were significant differences between the different notification groups, with remediation rates between 76.3 % and 33.9 %. Thus, the impact of the different factors considered in our experiment can be very high, especially in combination. We thus discuss them in more detail below.

Letters were effective, but expensive.

LETTERS ARE (ONCE AGAIN) EFFECTIVE As in the previous study, we showed letters to be the more effective medium, leading to an increase in remediation rate of between 3.9 and 17.9 percentage points (mean: 11.1) compared to an email with the same sender and framing. Nevertheless, letters also imply significant manual effort for contact data collection and financial costs for postage — our study required spending approximately 5000 € on domestic postage (for all three messages combined).

The legal sender was the most effective.

THE CHOICE OF MESSENGER MATTERS We observed significantly increased remediation rates when messages are sent by a legal research group (UNI-LAW). Compared to UNI-CS, remediation rates increased by 5.7 percentage points, a significant change (54 % to 59.7 %, $p < 0.05$, cf. Table 10). Interestingly, CITIZEN fell between these two groups, outperforming UNI-CS, and had no statistically significant difference to either of them. We discuss this curious result in more detail in Section 7.4.3.

The compliance framings significantly outperformed the baseline privacy framing.

COMPLIANCE FRAMING IS HIGHLY EFFECTIVE The largest observed effect came from changing the framing of the message. The PRIVACY framing was significantly outperformed by the GDPR and GDPR+FINE framings, with an almost 20 percentage-point remediation rate difference between CITIZEN and GDPR+FINE after the initial notification. In fact, all differences between the three framings for the initial message and the full timeframe were statistically significant, many highly so ($p < 0.0001$). This highlights the power of leveraging existing regulation in driving change, although we also found that these notifications were more likely to lead to negative reactions from recipients, including expressions of fear and complaints or legal threats.

Even with manual data collection, reachability remained problematic.

REACHABILITY REMAINS A CHALLENGE As in the previous study, we still observed a surprisingly high rate of messages being returned as undeliverable (3.5 % for letters, 5.8 % for emails). As the used data

source is legally required to be correct and up to date, this failure indicates that some of the website appear to not be maintained very well, which may be another factor in the presence of misconfigurations.

PERSISTENCE PAYS OFF Human error prevented us from conducting a detailed analysis of the effectiveness of our reminders. However, the groups that are unaffected by the error showed the reminders increasing overall remediation rates. The group with the largest effect from reminders showed 45.3 % of sites that were non-compliant five weeks after the initial message being remediated after the reminder. The overall remediation rates tell a similar story: prior to the reminder, 41.2 % of all notified operators had remediated, which increased to 56.6 % after the reminder. Although we had no control of websites that did not receive a reminder, these results indicate that reminders are likely effective.

Reminders were likely effective.

PROVIDING TOOLS IS HELPFUL We saw significant usage of our self-service tool, CheckGA: Assuming that all scans of the websites included in the study were conducted by notification recipients, 46.9 % of the notified system operators used the tool at least once, and 67.6 % of those that remediated successfully had used the tool. These numbers are likely slightly over-estimated, as the tool was shared by data protection experts and consultancies on blogs and social media, which led to an influx of users unaffiliated with any website from the study. However, given that only 3.1 % of sites in the control group were scanned, we can still assume that most of the scans of websites in the study were conducted by their respective operators.

The tool was widely used and perceived as helpful.

7.4.2 *Survey Results*

The results collected in the survey were frequently in line with the empirical results. However, we also found some interesting discrepancies, which we discuss in this section.

The survey responses frequently, but not always, matched empirical results.

NO SINGLE FACTOR INCREASES TRUST FOR EVERYONE Respondents cited many factors as increasing their trust in the message: everything from using the official letterhead of the sending institution down to the correct grammar and spelling and the signature at the bottom of the letter was noted as increasing trustworthiness. However, a minority of respondents stated that these factors actually *reduced* their trust in the message. This shows that no one message will be right for everyone, and designing an effective message always includes tradeoffs.

The same factor could have different effects on trust for different people.

RECIPIENTS DISTRUST UNSOLICITED MESSAGES The golden rules that IT security professionals keep teaching their colleagues to keep

them and their company safe — "don't open unsolicited messages", "don't click links from unknown senders", "when in doubt, don't respond", "if it is too good to be true, it probably isn't", etc. — actually work against IT security and compliance when it comes to notification campaigns. It is thus unsurprising that many recipients were initially wary of our message and sought verification through other channels. The CITIZEN – LETTER group in particular led to questions about the motivation of a private individual to spend money to notify the operators of websites about misconfigurations instead of simply writing an email. When given no more information, some recipients tended to assume commercial interests or bad intentions instead of altruism.

Distrust of unsolicited messages was common, especially if no clear motivation can be discerned.

PERCEPTION–ACTION RELATIONSHIP INCONCLUSIVE Interestingly, distrust of a message did not appear to necessarily imply inaction: While some senders and messages were rated as less trustworthy in the survey, this did not always translate into lower remediation rates. Most strikingly, messages sent by CITIZEN were deemed much less trustworthy than those sent by UNI-CS (cf. Figure 15), but still achieved statistically identical remediation rates ($p \approx 1$). Here, questioning the motives of the sender may have actually *increased* remediation, as some recipients may have seen it as a prelude to legal action. As we did not collect data on this theory, it remains only one of several possible explanations.

Distrust in a message did not necessarily lead to inaction.

A similar effect could be observed for EMAIL and LETTER, which had almost identical reported trust, but a difference of 11.2 percentage points in remediation rates at the end of the study. As discussed in the previous chapter, a possible explanation may be that the *a priori* trustworthiness attributed to letters was higher than that of emails, which may have made a "sketchy" letter still more trustworthy than a similar email. Anecdotally, this is supported by comments from some recipients that stated that they would not have trusted the message if it had been an email. A competing (or compounding) explanation may simply be that some emails were discarded as spam by automated systems, a risk which letters do not face. Spam filters may thus make up the difference in remediation rates. As we did not track the opening rates of emails, we have no way to quantify this effect.

Self-reported trust levels and remediation rates did not necessarily correlate.

Different mechanisms might be at play here.

RECIPIENTS DESIRE SUPPORT We received requests for individual support from 204 recipients (5.1 %). Supporting their remediation attempts with explanations and occasionally example code sometimes required multiple rounds of emails and significant time investment, but frequently led to successful remediations. Such a level of individualized support is infeasible for very large notification campaigns. However, the requests illustrate that detailed remediation instructions and automated tools are helpful and important, as they are a resource that recipients can be referred to. While we did not experimentally

Many recipients asked for support, highlighting the usefulness of providing a public tool.

validate this, it is safe to assume that providing a tool with extensive documentation about remediation reduced the amount of questions we received. However, the survey showed that such a tool is definitely appreciated by the recipients, with 87.2 % of respondents rating it as helpful.

OPERATORS LACK AWARENESS Interestingly, 19.5 % of respondents reported not knowing that Google Analytics was in use on their website. This indicates that they were not actually using the analytics data that was being collected, an impression that is reinforced by the surprisingly high number (36 % of remediating recipients) that removed Google Analytics in response to our message, instead of repairing the configuration.

Lack of awareness is a common problem.

This lack of awareness raises two interesting points: Firstly, our experience with sites hosted on Wordpress.com shows that some hosting services add their own tracking to their customers' websites without informing them (or giving them access to the data), which raises questions of liability — legally, the operator is responsible for any data collection on their website, even that added by their hosting provider that is outside of their own control. Secondly, it shows that a significant percentage of web tracking data appears to be collected but never looked at by the operators, for example because it was set up and then forgotten by a web designer, an explanation we received from multiple recipients. This tracking could thus be removed without negative impact on the website operator, thereby improving the overall privacy and compliance posture of the website.

Misconfigured services may have been set up by third parties that are no longer involved in maintaining the website.

7.4.3 *Comparison with Prior Work*

After discussing our results, we now consider how they relate to prior studies in this field.

We compare our results with prior work.

AWARENESS Many attempts to promote IT security and/or compliance rely on outreach and awareness to promote desireable behavior (e.g., installation of software updates, compliance to data protection legislation, ...). Without a doubt, awareness of a problem is a necessary precondition for remediation in most cases. However, our results indicate that awareness alone is not sufficient: 58.9 % of respondents in our survey reported having been aware that the IP anonymization exists, and 12.7 % even knew that they were not using it. These results are no outlier in the field of notifications: Durumeric *et al.* notified about the well-known and widely publicized *Heartbleed* issue. They found that *all* recipients had heard about Heartbleed before the notification, and many reported having already attempted to remediate, but had apparently not remediated all machines [34]. This trend was also observed by Li *et al.* who, notifying about a range of security misconfigurations,

Previous studies have shown that awareness alone is not sufficient for ensuring remediation.

found that 46 % of recipients had already been aware of the issue and 16 % had previously attempted remediation [68]. Çetin *et al.* also found that 40 % of notified operators of vulnerable DNS name servers had previously attempted to remediate [17].

Even when using manual data collection, bounce rates remain high.

BOUNCE RATES As discussed in Section 6.4, only one study outside this dissertation used manual address collection. Stock *et al.* performed a small-scale (N = 364) notification experiment with manual contact channels [98]. They found bounce rates of 0 and 26.8 % for emails and letters, respectively, but on a self-selected sample of recipients that had not reacted to previous fully-automated notifications. In our own prior study, described in the previous chapter, we observed bounce rates of 1.1 and 1.7 % for emails and letters, lower than the 5.8 and 3.5 % of this study. This discrepancy shows that even with the same address data collection methodology, there is still some variation in bounce rates on different samples. Of course, the delivery rates still exceed those of many previous studies using automated address collection, which often found bounce rates exceeding 50 % [17, 18, 99].

Guessing RFC 2142-compliant email addresses will become more important in the future.

Especially now that the WHOIS interface is no longer available for many TLDs, automatically-derived email addresses (e.g., following RFC 2142 [24] as used in previous studies [17, 98, 99]) will likely become a more important factor in future notifications. To gain a lower bound of the expected delivery success of such notifications on our sample, we briefly analyze our corpus of email addresses. We find that while some of our collected addresses match these common aliases — 41 % were of the form *info@domain.tld* and 0.8 % matched *{webmaster,hostmaster}@domain.tld* — many did not: 21.1 % used a different prefix, and 37.1 % were hosted at an entirely different domain. It is possible that functional standard alias addresses exist for more domains but were not listed in the imprint (our results only serve as a lower bound). However, in a study focused on determining good contacts for automated notifications, Soussi *et al.* found that the most-used generic alias was only available for approximately 68 % of domains in their sample [96]. This indicates that notifications may have to use multiple different addresses, which may in turn trigger more aggressive spam filtering. More work is needed to determine good address data sources for automated notifications.

Our manually-collected data implies that that these aliases aren't the desired contact channel for most operators.

We reproduced the improved performance of letters over emails, but obtained a smaller difference between them.

MEDIUM We once again observed increased remediation rates for recipients that received a letter compared to an email, with remediation rates of 60.3 and 49.1 % for letters and emails, respectively. These remediation rates are higher than in the study described in the previous chapter (59 and 40 %, respectively), although the difference between the groups is smaller. We have no definitive explanation for this discrepancy. It may be related to the different data sources used for the two studies, to fundamental differences within the populations

of websites with the respective issues (i.e., operators with information leaks on their websites may be fundamentally different from those with misconfigured Google Analytics installations), or to random chance.

MESSAGE SENDER The effect of the message sender has been the subject of three prior studies (cf. Section 2.3.2): Stock *et al.* compared messages that looked like they were sent by a human with those sent by an automated system [98], Zeng *et al.* sent part of their messages via the Google Search Console while others came from a university email account [117], and Çetin *et al.* used the identities of a private security researcher, a university, and a well-known organization in the field of malware research [18]. All three studies reported only small differences between the senders, which seems to conflict with our results that show statistically highly significant differences between the UNI-CS and UNI-LAW senders.

Previous studies found only small differences between different senders.

Previous studies assumed that differences between different senders would be explainable through differences in name recognition / sender reputation [18] (i.e., people may be more likely to trust a message sent by Google than one sent by a private individual they never heard of). However, these studies failed to show significant differences between the senders, and such an explanation also cannot explain why the CITIZEN group should be as effective (or even more effective than) the UNI-CS group in this study.

These studies theorized that differences in name recognition would lead to different remediation rates.

It may be that the differences in effectiveness in our study, particularly as it is dealing with an issue of compliance, can thus not be explained by name recognition or reputation, but instead by the perceived ability and willingness of the sender to *impose consequences for inaction*. While a computer science group at a university may be considered unlikely to pursue legal action against non-compliant website operators, a private individual with unclear motivation for reaching out is more of an unknown quantity. The same could be said for a legal research team, to which the recipients may additionally attribute a higher expertise in correctly determining questions of compliance. This explanation hearkens back to the concept of *leverage* from the transparency literature (cf. Section 2.4.2.2), according to which transparency can only be effective if the stakeholders have a plausible way to exert influence over the organization in question. The question of consequences also factors into the factor of framing and incentives, which we discuss next.

A competing explanation may be the perceived ability of the sender to impose negative consequences for inaction.

This matches existing theories on effective transparency regimes.

FRAMING AND INCENTIVES The use of different framings has been previously investigated by Zeng *et al.*, who notified about issues with TLS deployments. They used either a *user focus* (focusing on the impact of the issue on the user) or a *technical focus* (focusing on the technical background of the issue) [117]. They found no statistically significant

The use of different framings did not make a large difference in a previous study.

differences between the two framings. If we follow the theory that potential consequences are an important mechanism determining (in)action, this result would be unsurprising, as the major consequence of inaction (users will be unable to access the website) was described in both messages.

The idea of consequences also underlies other previous studies, which had success with explicit incentive mechanisms.

This theory also implicitly underpins other previous studies, which used incentives like browser warnings [69] or blocking the internet access of users suffering from Malware infections until they remediate [15, 16]. Çetin *et al.* also explicitly compared the effectiveness of such quarantine actions with those of email notifications, finding quarantines to be more effective [16]. These results suggest that direct and explicit incentives for remediation may be a powerful motivator for future notifications. Our results comparing the three different framings indicate that legal requirements and regulations can serve such a role, and preliminary work on using such arguments in more countries has already been conducted by Diop *et al.* [29]. It also matches the results of Çetin *et al.*, who reported that phishing and malware sites were cleaned up more quickly if they targeted banking credentials and thus posed a higher legal risk for the hosting company [18].

Establishing trust is a common issue in notification studies and remains an unsolved problem.

TRUST AND DISTRUST Distrust by message recipients [15, 16, 98, 117] and a desire to verify the authenticity of the message [15, 16] are commonly reported in previous studies. We also observed them in our previous study on information leaks (cf. Section 6.3.3). Our survey indicates that no one factor will increase the perceived trustworthiness of a message for every recipient. Conceptually, unsolicited notifications need to differentiate themselves from spam messages — however, spam and phishing messages have co-opted many of the same methods that notification senders may want to use to increase recipient trust, like using (forged) message senders with a presumably high reputation, using professionally-designed message templates, and giving incentives for (in)action. Thus, establishing trust remains a hard problem in practice.

One previous study found that providing a tool does not increase remediation rates.

SUPPORT TOOLS Automated tools that help message recipients to determine if their systems are affected by a misconfiguration and contain instructions for remediation have been frequently requested by recipients [17, 68, 69, 117]. However, only one study investigated the effect of providing or withholding such a tool: Çetin *et al.* investigated misconfigured DNS servers and provided a tool to verify if a server is vulnerable. They found no statistically significant difference in remediation rates between recipients that had been provided with the tool and those that had not [17].

We did not repeat this experiment in our study. However, our results (in particular our interactions with the recipients) indicate that providing a tool likely has other advantages: it simplifies supporting the

recipients in remediation, may reduce the amount of support requests (especially those asking for validation of their remediation), and (according to the survey) may increase the perceived trustworthiness of the message for some recipients. Thus, regardless of the effect on remediation rates, it may be in the best interest of notifiers to provide such a tool.

However, our results suggest that they may have different advantages.

REMINDERS Only two prior studies investigated the effectiveness of sending reminder messages, and came to inconclusive results: Stock *et al.* reported a small effect [99], while Li *et al.* saw none [68]. Our study comes to a different result: on average, reminders led 29.7 % of recipients that had not already remediated to do so. In some experimental groups, this number exceeds 40 % (cf. Table 11). Unlike Li *et al.* [68], but similar to Stock *et al.* [99], we did not have the presence of reminders as a varied experimental factor (i.e., there is no control group of sites that received an initial message, did not remediate, and did not receive a reminder). We thus cannot know what percentage of these recipients would have remediated without a reminder, although given the steep slopes in the survival rates after the reminders were sent, it is likely that they had a major effect on remediation rates. This discrepancy with prior studies remains unexplained, highlighting the need for more research in this area.

Previous studies on the effectiveness of reminder messages were inconclusive.

For us, reminders led to more remediations, but this part of the experiment did not have a control group.

SUMMARY Overall, we can confirm many results from previous studies: recipient distrust is an important issue, automated scanning tools are perceived as helpful, and awareness of an issue does not necessarily lead to remediation. We highlight the importance of incentives for remediation, as were already used by previous studies [15, 16, 117], and show that legislation may provide such incentives. When it comes to the effect of different senders and of reminders, our results diverge from previous research. This highlights that many questions in the field of effective large-scale notifications remain open for future researchers to answer.

We summarize our core takeaways.

7.5 LIMITATIONS

Our study has a number of limitations, both for internal and external validity, which we discuss here.

We discuss limitations to internal and external validity.

INTERNAL VALIDITY Four issues limit the internal validity of our study. The first of these issues consists of two types of self-selection in our dataset: address availability and survey responsiveness. Recipients for which we could only find one type of address (i.e., only email or only postal address) in the imprint were assigned to that group, i.e., they self-selected for a specific treatment in the message medium, but were spread equally between the groups for the other two varied

Our dataset had a certain amount of self-selection.

factors. This self-selection affects 87 email and 152 letter recipients, about 6 % of the total non-CONTROL dataset. The second type of self-selection is in survey participation, where it is likely that respondents skew towards higher trust in the messages, as those that distrust the messages may be less likely to respond to a survey.

We may falsely detect a website as having removed Google Analytics when they have added a cookie consent system.

The second issue that may affect validity is of a technical nature: our compliance scanner cannot automatically interact with cookie consent systems. Thus, any tracking that occurs only after a cookie consent banner is confirmed cannot be automatically detected, which may lead to false negatives if a non-compliant Google Analytics system is hidden behind a consent system. These cases would have been detected as a removal of Google Analytics from the website. We manually visited a subsample of 50 websites that our system detected to have removed Google Analytics and found no such false negatives (cf. Section 7.3.1.3), but did not check all websites.

The incorrect reminder messages have an unknown impact on the study.

The third issue is the incorrect set of reminders we sent to part of the LETTER – UNI-LAW group. We cannot quantify the effect of this mistake and have thus refrained from a detailed analysis of the reminders. However, the major trends were already visible before the reminders, and the unaffected groups give many indications that the reminders were generally effective. It is also likely that some recipients received messages from more than one experimental group, as a single web design agency may have been hired by multiple website owners that were assigned to different experimental groups. We received a small number of indications that such leaks between the groups took place. In these cases, observer effects may impact the results, as the recipients may suspect that they are part of a scientific study and thus change their behavior.

A small number of recipients may have received messages from more than one group.

Different mail server configurations may have had an unknown impact on the deliverability of messages.

The fourth and final limitation of the internal validity comes from the fact that we used three different senders, which required the use of three different mail servers. As these servers did not have identical configurations (for example, the UNI-LAW server was lacking Sender Policy Framework (SPF) and DomainKeys Identified Mail (DKIM) records) and reputations, this may have led to differences in message delivery due to spam filters — a problem that previous studies also encountered and were unable to control for [18, 117]. We are hopeful that these differences did not make a large difference for the notifications, as we observed similar bounce rates for the different servers, and the rate at which CheckGA was accessed was highest for UNI-LAW, the server with the most incomplete configuration.

Recipients were spread over many different mail servers.

To gain more insight into the diversity of recipient email servers, we conducted a small analysis a few months after the end of the study. The 1337 recipient email addresses use 516 different second-level domains. The average mail provider handles 2.5 of these addresses (median: 1), and even the most common providers like Outlook.com and Google only handled 108 and 70 addresses, respectively. As the

messages were sent over a timeframe of 5 days using three different senders, we hope that the impact of spam filtering will be limited.

EXTERNAL VALIDITY There are two major limitations to the external validity of our study. Firstly, our sample is likely not representative for the overall population of websites, either in Germany or worldwide. However, the more important limitation is our focus on German websites. It is unknown if these results can be transferred to other countries, with their own laws and cultures. This limitation is a direct consequence of our experimental focus on compliance as a factor: compliance is, by its very nature, rooted in local law, as are the expectations of website operators about the likelihood of being targeted by enforcement action. Any message making use of a compliance argument thus needs to consider the specifics of the local situation.

The results for German websites may not be transferable to other cultures and jurisdictions.

Naturally, the effort of such tailored notifications is higher the more countries and legal systems are involved. Any notification campaign attempting this at larger scales would be well-advised to partner with local organizations that are familiar with local laws and practices. Several prior studies have already made use of such local organizations [60, 68, 69].

We recommend partnering with local organizations for tailored notifications on a per-country basis.

7.6 CONCLUSION

In this chapter, we described the results[17] of a notification experiment involving 4594 recipients that operate 4754 distinct websites that used Google Analytics without the legally mandated anonymization function enabled. We observed an overall high effectiveness of our notifications: the overall remediation rate was 56.6 % for notified sites, compared to 9.2 % for the control group. The results highlight the importance of using a good framing of the issue, and replicate the increased notification effectiveness obtained from the use of letters instead of emails that we already found in the previous chapter.

We described our third notification study, which had a large effect on the notified websites.

Our results show that the use of legal obligations as an argument for remediation can increase remediation rates by almost 20 percentage points compared to the worst-performing argument. Interestingly, this effect was obtained despite the fact that none of the notification senders actually had any legal power to directly impose fines for non-compliance. These results highlight the potential benefits of collaborating with legal experts on notification campaigns — a collaboration that also proved invaluable in deflecting complaints and legal threats by a small number of disgruntled notification recipients.

Legal obligations have proven to be a potent argument.

We further extended our knowledge of the factors that influence recipients' trust in notification through a survey with 477 respondents, which highlighted a number of formal and content-related factors that

We also collected information directly from the recipients using a survey.

17 The used messages, code and data underlying the study can be found on Zenodo: https://zenodo.org/record/4075131 (last accessed 2021-07-15).

influence the perceived trustworthiness of a message. The survey also showed that almost a fifth of respondents were unaware that Google Analytics was active on their website. This indicates that a significant fraction of tracking data is never used — an interpretation that is also supported by the 36 % of remediating recipients that chose to remove Google Analytics completely instead of activating the anonymization function, and some even chose to take their website offline. Thus, a secondary effect of our notifications was to motivate recipients to disable unmaintained systems and services on their website, which improves the overall security and privacy posture of the web.

Part IV

CONCLUSIONS

We synthesize the practical experience of conducting our three notification studies into a set of recommendations for future studies in this area. Afterwards, we conclude with a discussion of the main results of this dissertation.

8

LESSONS FOR NOTIFICATION CAMPAIGNS

While conducting the studies described in the previous chapters, we learned many important lessons about the do's and dont's of large-scale internet scanning and notifications.[1] In order to support future research in this area, we document these lessons here to allow others to benefit from our experience and avoid making the same mistakes. We begin by discussing the technical issues of medium- to large-scale internet scanning before moving on to the challenges of sending large amounts of notifications. We then discuss best practices for interacting with and supporting notification recipients before closing with the topic of follow-up messages like reminders. However, first of all, we will highlight a central concern that applies to all aspects of a notification study: research ethics.

We distill best practices and lessons for future notification campaigns.

RESEARCH ETHICS Notification studies are likely to be classified as human-subject research by many institutions. This is especially true if the study design includes deceiving the notification recipients about the intention of the messages (e.g., by not disclosing that they are part of a study to avoid observer effects). The study design should include provisions on lifting this deception after the study has ended (which can be combined with a survey of recipients, if desired), and also plan for a notification to members of the control group to allow them to remediate as well. Researchers should familiarize themselves with the processes for ethical review and seek approval as early as possible to avoid delays caused by missing ethics approval. For further reading on ethical aspects of large-scale measurements and online data research, we refer to the works by Vitak *et al.* [111]. The companion paper to this chapter [72] also contains a more detailed discussion.

Human subject research frequently requires approval by an ethics board.

8.1 INTERNET SCANNING

Many notification studies begin with a data collection phase, which frequently requires large-scale scanning of internet system to detect the issue(s) at the heart of the study. We first discuss the consideration for the design of a data collection stack before proceeding to some lessons we learned for the data collection process itself.

Notification studies begin with a data collection phase.

We explicitly *do not* consider the choice of the initial dataset of systems to be scanned, for which we have no special experience.

Selecting a list of scannable systems is out of scope for this chapter.

1 This chapter contains references to both Open Source and commercial services. We included services based on familiarity and did not receive any compensation, financial or otherwise, for including them.

However, we do point out that the most common data source for web scans, the Alexa Top Million websites, has been criticized as unstable and potentially unrepresentative [64, 94], and an alternative exists that is, according to its authors, better suited for some use cases [64].

8.1.1 *System Design*

Large-scale scans require a detection software and a test harness.

Large-scale scanning requires the development of two different software components: the actual detection software (if no pre-existing software can be used for this) and a test harness that runs the detection software on a large list of targets, saves the results in a database, and gives insight into the current state of the scans.

The choice of detection software depends on the issue at hand.

CHOOSING A DETECTION SOFTWARE Depending on what kind of misconfiguration should be detected, different types of scanners are required. For example, to simply test if a file at a specific path exists and contains a certain string, a simple script can be sufficient (cf. Section 6.1.3), while an issue that requires the evaluation of JavaScript may need a complete browser environment to be detected (cf. Section 7.1.3 — two ready-made systems for this purpose are OpenWPM [38] and PrivacyScanner [87], the latter of which was developed by Pridöhl *et al.* as part of the PrivacyScore project described in Chapter 3). The use of an instrumented browser is a significantly more complex and resource-intensive process, but is often the only way a modern website can be automatically evaluated, as they frequently rely on JavaScript for large parts of their functionality. Other studies may be able to use existing tools, in which case they should make themselves intimately familiar with the use and limitations of the tool.

Many types of errors can occur during the scanning process and need to be handled.

KNOWING THE ERROR CLASSES In both cases, the researcher needs to consider how the detection software behaves in a variety of possible error cases. What happens if the network connection is severed in the middle of a test? If the target is unreachable? If the machine running the scans runs out of disk space? Does the tool stop with an error? Does it have a retry counter? Does it freeze and has to be stopped manually? The test harness should be adapted to handle and log these cases. Particular attention should be paid to how the system behaves if multiple scanners are run in parallel on the same machine, as some systems may unintentionally begin sharing state between instances (browser profiles, temporary files, ...).

INCLUDE TEST CASES We recommend operating one or more test systems where the correct result of the scan is known.[2] These systems

2 An example of a public system of this type is `https://badssl.com/`, which hosts test cases for many types of broken TLS configurations that should be rejected by clients. Last accessed 2021-01-12.

can be used during development to test the system. However, they can also be included in the later regular scans as test cases that can indicate if a previously working system has stopped working. Additionally, the scans could also include a *dead man's switch* — a target that expects to be accessed once per scan interval, and will send a notification if it is not, as this may indicate that the scans have stopped working. Such a service is offered by several companies like Healthchecks.io[3], Dead Man's Snitch[4], or PushMon[5], which often include a free service tier that should be sufficient for such a study.

Some systems with known-good or known-bad configurations should be included as test cases.

PREPARING FOR UNKNOWN ERRORS It is almost inevitable that new error classes will appear once the system begins large-scale scans. It is thus a good practice to enforce specific expectations of the behavior of the scanning system to be able to detect these new errors. Among other things, this includes the output format and the expected duration of a scan (to detect and handle deadlocks / livelocks). Deviating behavior should be reported (use of a tool like *Sentry*[6] is recommended to collect errors and uncaught exceptions during development and in production), with affected results being marked in the database for later inspection and debugging.

Not all error types will be known in advance.

PLAN FOR SCHEDULED AND ONE-OFF SCANS Many studies require (at least) two phases of scanning: An initial one-off scan of a very large dataset, followed by regular scans of a smaller set of systems that are part of the study. Both of these use cases must be supported by the scanning infrastructure.

Both one-off and scheduled scans should be supported.

DISTRIBUTING THE SCANS While not strictly necessary, running redundant copies of the scanning infrastructure on different machines and different networks can increase confidence in the results. In one of our experiments, we encountered problems with a misconfigured IDS appliance that led to one of the four scheduled daily scans reproducibly failing for most websites due to timeouts. Such mistakes, which would usually lead to missing data, can be compensated if the same data is collected from more than one point of origin at the same time. Checking for disagreements between the results of the different scanners can also be a good method to detect further errors in the scanning stack.

Researchers should consider using a distributed scan infrastructure.

3 See https://healthchecks.io, last accessed 2021-01-11. Transparency note: I received a free upgrade to the *Business* tier under the Open Source support program of the service (https://healthchecks.io/faq/#free-for-open-source, last accessed 2021-01-11). Healthchecks.io is Open Source and can be self-hosted or used as a commercial cloud service.

4 See https://deadmanssnitch.com/, last accessed 2021-01-11. Dead Man's Snitch is a commercial service.

5 See https://pushmon.com, last accessed 2021-01-11. PushMon is a commercial service.

6 See https://sentry.io, last accessed 2021-01-11. Sentry is Open Source and can be self-hosted or used as a commercial cloud service.

Scans should be performed frequently.

SCAN OFTEN If feasible, we recommend scanning each target more than once per day. Firstly, doing so gives a more detailed and fine-grained view of when a website remediated. Secondly, having more results is helpful when a scanner may occasionally return incorrect results (e.g., due to connection issues) — in these cases, having more than one scan per day allows the researcher to fall back on the other scans to determine the state of the target system. Finally, some systems may show cyclic behavior that is only revealed when considering multiple scans per day. For example, in our Google Analytics study, one website switched between a day and night version depending on the time of day, with the misconfiguration being present on only one of the versions (cf. Section 7.2.4).

The system should be protected against data loss.

BACKUPS It goes without saying that automated, regular, off-site backups are an important part of any data collection system. This is doubly true if the scanner is operated on third-party machines which may be deactivated without prior notice if the provider receives abuse notifications or has automated filtering systems of their own that may interpret the scans as outgoing attacks.

If the scanner should identify itself as part of a research project depends on the details of the study.

IDENTIFYING THE SCANNER Researchers should carefully consider if and how the scanner should identify itself as part of a research project. Potential methods include the use of a custom user agent (for website scans) and the use of a custom reverse DNS entry for the IP(s) of the scanning machines. This may reduce the amount of abuse notifications from recipients that interpret the scans as an attack. However, it may also introduce observer effects (i.e., notification recipients may be able to identify that they are part of a scientific study, which may change their behavior). Which of these two possibilities is more important depends on the study design and invasiveness of the scans, and is left up to the researcher to decide.

8.1.2 *Data Collection*

We discuss what data to collect.

The question of which data should be collected is important enough to warrant its own section. We begin with some general remarks before proceeding to more specific data collection recommendations and considerations that apply to the area of websites, as these were the subject of our own studies.

Seemingly unimportant data points may become relevant later and should be collected.

WHEN IN DOUBT, COLLECT THE DATA In an ideal world, a researcher would know which data they will need to collect for their evaluations before writing the first line of code. However, experience has shown that during the evaluation of the data, new questions may appear that require new data points. If this data has not been collected, such questions remain impossible to answer. We thus argue

that researchers should follow a strategy of "when in doubt, collect it." This includes the identity of the scanning machine (if more than one is used), the IP address of the scanned machine, response status codes (where applicable), the complete output of any tool that is being used, extensive log files including timestamps for each scan, and (where applicable) the data the results have been derived from (like the source code of scanned websites, network traces, etc.). For additional best practices for scientific code, we refer to the article by Benureau and Rougier [9]. Best practices for data collection are given by Bajpai *et al.* [6] (on general networking research), Cui *et al.* [25] (on large-scale scanning), and Durumeric *et al.* [33] (on port scans).

However, technical, ethical and legal requirements need to be taken into account.

Of course, in practice the benefits of saving the data need to be weighted against the cost in terms of complexity and storage space, and in some cases even legal considerations (e.g., when looking for database backups that are unintentionally published online, avoid downloading and saving the entire database, and download only the first few thousand bytes that allow validating the file format, as described in Section 6.1.3). Ethical aspects of data collection are discussed by Vitak *et al.* [111]. Marcia Hoffmann wrote about the legal aspects of scanning for Rapid7[7].

There is frequently some time between the initial scans and the beginning of the study.

BEGIN DATA COLLECTION EARLY At some point, the researcher will need to begin conducting regular scans of all websites in the dataset to monitor the effects of the notifications. Frequently, there are a few days or even weeks between the first scan (which determines which websites are in the final dataset of notified websites) and the day the first notifications are sent.

Beginning the data collection early can be advantageous.

We recommend beginning the regular data collection as early as possible, even before the notifications are sent. Having one or two weeks of scans before the notifications are sent serves two purposes: firstly, it can help increase confidence in the stability and correctness of the data collection system, as it can reveal previously-undiscovered bugs before they can impact the more important measurements after the notifications are sent. Secondly, if the dataset shows that notified and unnotified websites show the same behavior for the weeks preceding the first notification, it can serve as a further proof for the representativeness of the control group. If the groups already show significantly different behavior at this point, it may be an indicator that the strategy behind the group assignment needs to be reconsidered.

Domains can forward to other websites, and these forwards can change over time.

WEB-SPECIFIC ISSUE 1: FORWARDING When scanning websites, researchers should be mindful of HTTP forwards, which may change which website is scanned. These forwards can result in two classes of issues: either two domains in the dataset can forward to the same

7 See https://www.rapid7.com/blog/post/2013/10/30/legal-considerations-for-widespread-scanning/, last accessed 2021-05-17.

website (leading to two seemingly-distinct scans giving the same result), or a single domain may forward to more than one domain over time, leading to different results over time. The latter is more common than might initially be expected, due to the existence of so-called *traffic services* (cf. Section 7.2.7).

A scanning system needs to implement support for following such redirects, if this is desired.

This issue needs to be addressed on three levels. First and most importantly, the scanning system needs to actually support forwards. This is trivially achieved when using an instrumented browser for the scans. Less sophisticated systems like downloading scripts will follow some types of redirects (HTTP 3XX redirects[8]) but not others (those triggered by JavaScript's `window.location` property[9] or HTML `meta` tags[10]).

Researchers need to decide if and how to incorporate these forwards into the dataset.

Second, the researcher needs to decide if their scanning system should keep using the domain from the initial dataset (e.g., the Alexa Top Million or Tranco toplist [64]), or follow all forwards once, save the final URL, and then keep using this URL for all future scans. The former is more representative of a user that keeps visiting a specific domain by typing it in by hand (and thereby following all redirects every time they visit the website), while the latter reduces the impact of traffic services and thus leads to more consistent results, at the cost that a website relaunch may invalidate the URL and lead to scans of error messages instead of the newly restructured homepage.

The final URL after all forwards should be logged in the database.

Third, in both cases, the scanning system should log the final URL after following all redirects and associate it with the obtained results in the database. Depending on what data is collected for the study, the researcher also needs to consider how to handle data collected from intermediate pages that may have triggered more network requests than a simple redirect. For example, when conducting a study of tracking and advertising code on websites, a website may contain a message "this website has moved, you will be redirected in 5 seconds" which will redirect the visitor after a specific timeframe, but already contain tracking or advertising code. Should this code count towards the results or not? Depending on the answer, the data collection process needs to be adapted.

Different domains can forward to the same target, biasing the dataset.

The case of two websites forwarding to the same final domain also needs to be considered in the evaluation, as it may otherwise give undue weight to the operator of a single website. Accordingly,

8 See https://en.wikipedia.org/wiki/List_of_HTTP_status_codes#3xx_redirection for a list of 3XX status codes, last accessed 2021-01-11.

9 See https://developer.mozilla.org/en-US/docs/Web/API/Window/location for details on the `window.location` property and https://www.w3schools.com/howto/howto_js_redirect_webpage.asp for instructions on redirects using it. Last accessed 2021-01-11.

10 See https://developer.mozilla.org/en-US/docs/Web/HTML/Element/meta for details on the `meta` tag and https://www.w3docs.com/snippets/html/how-to-redirect-a-web-page-in-html.html for instructions on redirects using it. Last accessed 2021-01-11.

a deduplication step should take place, ideally before sending the notifications but at the latest during the evaluation of the results.

WEB-SPECIFIC ISSUE 2: DYNAMIC AND HIDDEN CONTENT Specific features of a website may be gated behind interactive content. The most prominent of these are the prevalent *cookie consent* banners, which may hide the presence of tracking and advertising code until the message has been confirmed. If such data is relevant to the study, cookie consent banners present a major challenge, as there is no standardized method for interacting with them that can be expected to be compatible with all websites. Researchers should be aware of this limitation of automated scans.

Dynamic content like JavaScript may hide specific functionality from automated scans.

WEB-SPECIFIC RECOMMENDATION: USE THE INTERNET ARCHIVE In some cases, it may be desireable to be able to "go back in time" to check what a website looked like at a specific point in time. For example, during the Google Analytics study, several recipients claimed that they were compliant and had never been non-compliant. In such cases, having an impartial source that contains the state of the website in question at a specific point in time can be valuable, if only to ensure that the detection systems did indeed work correctly and the system operator is incorrect.

Sometimes, determining what a website looked like at a specific point in time can be important.

The Internet Archive is a non-profit organization that maintains the *Wayback Machine*, a system that collects snapshots of websites over time and allows anyone to inspect older versions of a website. If the issue in question can be detected from a manual inspection of the source code of a website, it may be advisable to submit each scanned website to the Wayback Machine with each scan to have an archived copy of the website at that point in time. Websites can be submitted for inclusion into the Wayback Machine with a simple GET request[11]. Alternatively, researchers can run their own website collection system[12].

This can be achieved using the Wayback Machine of the Internet Achive.

8.2 NOTIFICATION PREPARATION AND DELIVERY

Once a set of affected system has been identified and the infrastructure for regular scans is in place, the next step is the preparation of the notification messages. We first discuss preparations for sending messages, and then highlight a few failure modes of the deliveries that make attributing undeliverable messages challenging.

We now consider the issues inherent in sending notification messages.

11 There does not seem to be official documentation on this, but you can simply send a GET request to `https://web.archive.org/save/[some-url]` and the URL will be added to the Internet Archive. Please set a HTTP user agent that identifies who you are and has a way to contact you as a courtesy to the operators of the Internet Archive.

12 See https://github.com/webrecorder/pywb for an Open Source project that mirrors the functionality of the Wayback Machine. Last accessed 2021-01-11.

8.2.1 *Preparing Messages*

Collecting address data is out of scope for this chapter.

We will not discuss the collection of address data, as this can vary depending on the notification campaign. Instead, we assume that addresses have been manually or automatically collected, and discuss the question of preparing and delivering the notifications. For a survey of possible automated channels for address collection, we refer to Soussi *et al.* [96].

A high level of automation is recommended.

THE GOLDEN RULE: AUTOMATE EVERYTHING If the planned notification campaign includes over 50 recipients, we highly recommend automating as many steps of the process as possible. What exactly is automated in what way depends greatly on the goals and varied factors of the notification campaign. It can be as simple as using a plugin for a mail client to generate many messages from the same pattern or as complicated as a custom web-based system that can send emails and receive responses.

Even with automation, human errors are still possible.

As an example: Our final notification study, featuring the Google Analytics misconfiguration, used a Thunderbird plugin[13] to generate the emails from a template, and a custom Python script that filled and compiled a LaTeX template for the letters. A mistake in one of the templates, introduced while debugging, led to incorrect reminder messages being sent for one group. This highlights that even with a high degree of automation, mistakes still happen, and manual sanity-checking of the output is necessary. For emails sent from an automated system, a test server that accepts emails and shows them in a web frontend may prove helpful. Examples include MailHog[14] and MailSlurper[15].

The configuration of the mail server can influence delivery rates.

EMAILS: CHECK THE BASICS When sending email notifications, researchers should make themselves familiar with the typical email security and authentication methods like DKIM and SPF, and ensure that the email server they are using has implemented them. Anecdotal experience shows that even large universities may not have properly configured their email servers, which may increase the chance that a message is dropped as spam by the recipients' mail server.

Some services send notifications if a mail server is listed on spam lists.

If the researchers have control over the email server they are using, they may also want to sign up for email server reputation monitoring systems like the Microsoft Junk Mail Reporting Program[16] and regu-

13 See https://addons.thunderbird.net/addon/mail-merge, last accessed 2021-01-11. Mail Merge is Open Source.

14 See https://github.com/mailhog/MailHog, last accessed 2021-01-11. MailHog is Open Source and self-hosted.

15 See https://mailslurper.com/, last accessed 2021-01-11. MailSlurper is Open Source and self-hosted.

16 See https://mail.live.com/mail/services.aspx, last accessed 2021-01-11.

larly checking spam blocklists like SpamCop[17] and Spamhaus[18]. This can give an indication if the mails are being filtered as spam or not.

EMAILS: PATIENCE IS A VIRTUE We recommend spacing out the sending of emails instead of sending them all in one big burst. Firstly, some mail servers will enforce such rate limiting anyway, and so error messages can be avoided by proactively slowing down. Secondly, while many spam filters work based on proprietary algorithms, common wisdom holds that big bursts of almost-identical emails raise red flags for spam filters. Slowing the messages down a little (by sending one every 30 seconds) may help alleviate such filtering. It also has another advantage when it comes to delivery failures, which we discuss in the next subsection.

Sending emails spread over a longer timeframe has several advantages.

8.2.2 *Delivery Failure Modes*

Often, researchers are interested in which messages were delivered successfully and which were undeliverable (although spam filters that discard without notice may make such statistics unreliable for emails). Collecting this data can be more challenging than may be assumed at first glance. We enumerate some of the challenges and potential workarounds here.

Emails have many delivery failure modes.

DELIVERY DELAYS Some email servers may attempt to deliver a message, fail, and then retry the delivery for a few days. Some email providers, most notably Google, will retry for up to 3 days and give daily updates about the continuing delay in delivery. This means that even messages that were not returned as undeliverable may be silently undelivered for a day or more, although email servers should give a notification once they stop further delivery attempts.

Delivery may be retried for multiple days before a failure is generated.

UNDELIVERABLE MESSAGES (BOUNCES) If an email is undeliverable, most email servers will notify the sender. This notification can either come from the sending email server (if the recipients' email server cannot be found) or from the receiving server (if the server exists, but the email account does not, or it has exceeded the quota). These responses (commonly called *bounces*) do not follow any standardized form. They can be as simple as a response to the message that includes a full quote of the original messages. However, they can also change many things about the original messages, including the subject line and the email address.

Bounce messages do not follow any specific standard.

In the worst case, a bounce messages may not be attributable based on email address or subject line. In these cases, spacing emails to be sent every 30 seconds can help narrow down the possibilities for

17 See https://www.spamcop.net/bl.shtml, last accessed 2021-01-11.
18 See https://www.spamhaus.org/lookup/, last accessed 2021-01-11.

In some cases, bounces may not be attributable to a message automatically.

who sent the bounce based on timestamps. While they can be helpful, researchers should not assume that any automated heuristic will be 100 % successful in attributing bounces to their originating message. Any automated system for this purpose should leave room for manual intervention by the researcher. Alternatively, an automated system could use custom email addresses for each sender. However, such automatically-generated email addresses may impact the spam scores and/or perceived trustworthiness of the messages and the tradeoffs should be carefully considered.

Autoreplies may direct the sender to a different contact channel.

NON-BOUNCE AUTOREPLIES Another challenge comes in the form of automated replies that either state that a person no longer works at a company and the message will not be forwarded, or that redirect to a web-based form. Researchers should decide in advance how they want to handle these cases consistently: will they invest manual effort to fill out forms or look for alternative contact details, or will they count these messages as undelivered? What would this mean for the results?

8.3 SUPPORT TOOL

We discuss best practices for self-service tools.

In our experience, it is very helpful to provide a tool that allows the recipients to validate if their remediation attempts were successful. Although the effect on remediation is reportedly low [17], it will likely reduce the amount of questions, and definitely make answering them easier. We briefly discuss some of the lessons we learned from operating such a tool for our two studies (cf. Chapter 6 and 7).

The tool needs to be clearly understandable by non-expert users.

MAKE THE TOOL USEFUL AND USABLE This may seem like an obvious point, but it bears repeating: the tool should have the look and feel of a (semi-)professional system. It should clearly state what it does, how to use it, and what the results mean (otherwise it will cause more support requests than it solves). If practical, it should contain clear information on how the issue can be remediated. It is worth investing time into these instructions — in the Google Analytics study, we found some users of the tool had copy-pasted the remediation code but not entered their own Google Analytics ID, leaving it at the placeholder value (UA-XXXXX-Y) and rendering their analytics system non-functional. Such mistakes may be preventable through better documentation. However, a previous study by Çetin *et al.* also reported that recipients had trouble understanding how to use the tool or used it incorrectly, even though clear instructions were prominently presented on the tool page [17, p. 7]. Thus, the tool should also offer helpful error messages when used incorrectly, and ideally try to predict the most common types of usage mistakes and offer specific guidance if these are detected.

Unclear documentation can lead to mistakes by the recipients.

CONSIDER BUILDING A FLEXIBLE TOOL Some recipients may want to scan other sites under their own control. If knowledge of a misconfiguration on a website does not result in immediate danger, researchers should consider allowing the tool to be used on arbitrary websites, instead of using custom links for each recipient that limit them to their own website. This allows them to collect information on other websites operated by the same party, which may be an interesting dataset in its own right. It also reduces the amount of messages asking the researchers to scan additional websites by the same operator, a phenomenon we observed in our second notification study (cf. Section 6.3.3).

Consider allowing the tool to scan arbitrary sites.

MAKE THE TOOL FINDABLE Some recipients (rightly) distrust links from unsolicited messages. In these cases, the participants may instead attempt to find the tool using search engines, and click through to it from there. Researchers should ensure that the tool can be found by the popular search engines by directly submitting it to their index[19] and ideally also linking to it from a university website, if this is compatible with the purpose of the tool and the story told in the notification. Of course, all of this is only possible if the tool can be used on arbitrary websites and does not require a custom link, as discussed above.

Ensuring the tool can be found with common search engines can increase trustworthiness.

COLLECT THE DATA When offering a tool, it should be used to collect data as well. We recommend logging which systems were scanned when, by whom, and what the results were. This will allow analyses such as: how many times did recipients use the tool before they remediated? How long from the first to the last scan? What other sites did they scan (if this is allowed by the tool)? Researchers should think about different states the misconfiguration may enter over time: is it possible that the first remediation attempt changes the misconfiguration to a different variant of the same issue, and if so, can this be detected? All of this can be interesting data to collect. This data should be easily distinguishable from the automated, scheduled scans to avoid mistakes in the later analysis.

A tool can be a valuable source of research data and should be treated as such.

BE AWARE OF ALTERNATIVES TO YOUR TOOL There may be more than one tool seeking to detect the issue in question. In our Google Analytics study, we found that a different tool existed that purported to detect the same issue, but was in fact returning faulty data and missing certain instances of the misconfiguration. This led to questions from notification recipients. Researchers should be aware of any alternatives to their own tools, and ideally be familiar with the error classes of these tools, if any are known.

Other tools may perform the same checks and be used by recipients.

19 Submitting websites to Google and Bing now requires an account with their *Webmaster tools* systems, see `https://developers.google.com/search/docs/advanced/crawling/ask-google-to-recrawl` and `https://www.bing.com/webmasters/homepage`, respectively. Last accessed 2021-01-12.

Hosting a second, private instance of the tool can be helpful for the researchers themselves.

HAVE A PRIVATE INSTANCE OF THE TOOL Naturally, having a tool that can detect the misconfiguration can also be helpful for the researcher. It can be used to check the current status of a misconfiguration while answering a message from the operator of that system, or allow easy comparison of the results of different tools. Any scans conducted by the researchers should be kept separate from the data collected from notification recipients to avoid contaminating the research data. The easiest way to achieve this is to host a copy of the tool under a different address that is only known to the researchers, which either has logging completely disabled or logs to a different database that will not become part of the analysis.

8.4 INTERACTING WITH RECIPIENTS

We discuss best practices when interacting with recipients.

Our experience shows that many notification recipients will reach out with a variety of intentions and questions, using a number of different channels. Preparing to receive and respond to these messages is an important step in making this part of the process as painless as possible. Nevertheless, if researchers are planning to respond to incoming messages (which will likely have a large impact on remediation rates), they should expect to invest significant time and effort into this part of the study.

It is worth investing in tools to aid in interactions with recipients.

We strongly recommend building tools to make interactions with recipients easier. In particular, when answering a phone call, it can be invaluable to have a quick tool that allows searching for a half-understood domain name using fuzzy searching (for example, trigram distance to the domain name and/or name of the recipient), especially if the study design requires telling different stories to different groups of recipients. The tool can also contain a link that easily starts a scan of the target system using a private, researcher-only instance of the self-service tool, so the researchers can more easily see what the current status is, and act accordingly. The tool can also be used to track incoming and sent messages (automatically or manually) if they should be evaluated in more detail later.

8.4.1 *Communication Channels*

Recipients will contact researchers via different channels.

While most questions will be sent via email, some recipients will prefer other communication channels, either because they believe them to be more effective or because they distrust emails and want to verify the messages using a different medium.

Expect to be called on the phone if the number is listed somewhere.

PHONE CALLS The most popular alternative medium is the phone. Our second study (cf. Chapter 6) showed that not providing a phone number in the notification does not ensure that recipients won't call. Instead, they either searched for the sender on the website of the

university and used the phone number provided there, called the secretary of the research group (as this was the most prominently listed phone number on the website), or even called the central switchboard of the university and asked to be connected.

This experience taught us two lessons for the third study: firstly, we warned all members of the research group that they may be getting phone calls and who to forward these to. Secondly, we included a phone number in the notification to reduce the impact on other people in the group, with the rationale that if people were going to call either way, they should at least directly reach the right person. This proved to be mostly effective, although some still called other phone numbers because they mistrusted the message. Giving specific time windows for calls in the notification was only moderately successful, as about half of the calls came outside the stated hours. Still, when giving a phone number, we still recommend giving a time window for calls.

Colleagues should be informed about the study in advance.

ALTERNATIVE EMAIL ADDRESSES Other recipients may prefer to send an email, but distrust the sending email address, especially if it is a function-specific account (i.e., notification@group.university.tld instead of name@group.university.tld). These may seek out an alternative email address, frequently falling back on the email listed on the university homepage. In some cases, they also found a private email address from a blog or other sources. This also highlights that anyone identified by name in the notifications should be aware of the information that will be found if recipients enter their name into a search engine, and consider how this information would change what recipients think about the notification.

Responses may be sent to different email addresses.

While such initiative and distrust of unknown senders is to be commended from an IT security perspective, it also implies that researchers should never assume that all relevant email responses will come to the intended email address (i.e., that used to send the email or listed in the reply-to field). This has two implications. Firstly, it means that identifying the sender of a response based on their reply to a recipient-specific email address will not always work, and secondly, any tool that is used to manage the communication with recipients should allow for the manual addition of incoming emails by the researcher, in case they use an email address not monitored by the tool.

This can lead to replies being missed by automated data collection systems.

ALTERNATIVE RECIPIENTS Some recipients take things a step further and intentionally contact a different person at the same institution. Examples include the head of the research group, the secretary, the general contact of the department, or the press office of the university. From there, they were then usually forwarded to the researcher managing the notification campaign, which was frequently only possible because this researcher was identified by name and department in

Some recipients will intentionally contact other people at the same institution.

the notification message. This once again highlights that notification campaigns can cause work for people outside the immediate study team, and that such contacts should be warned in advance that they may receive calls or emails about this study.

Other channels, like social media, may also be used.

LESS-USED CHANNELS Finally, we also observed a few atypical channels that were only used by very few recipients. These included direct messages on Twitter and professional social networks like LinkedIn or Xing. Anyone mentioned by name in a notification should regularly check any accounts they maintain under their real name at such networks.[20]

8.4.2 *Requests for Help*

Recipients may also request help in remediating the issue.

Aside from validating the authenticity of the message, requests for help will likely be common. We briefly discuss a few considerations for these cases.

Researchers should decide in advance whether to provide this help, and do it consistently.

HAVE A HELP POLICY The first question is if the researchers should attempt to help recipients or not. Helping them will increase remediation rates, while not helping them will likely lead to resentment. This question should also be considered under a scientific lense: would helping the recipients distort the results (for example, because one group has better access to help than another)? Or is it unethical *not* to help, due to the high danger or complexity of the issue? Researchers should decide on a policy for this before sending the first notification.

Recipients may wrongly think that they resolved the problem.

HAVE A CORRECTION POLICY A sub-problem of this issue may be that some recipients may reply to thank the researchers and claim that they have remediated successfully while the scans show the issue as non-remediated. Again, researchers should consider if they want to correct these faulty claims or not, considering the potential biases they may introduce weighed against the risk of non-remediation in the process. If they do decide to correct the claims of these recipients, they should be prepared to offer proof of their assertions and provide more detailed instructions on remediation.

Recipients may also ask for help with unrelated problems.

BE PREPARED FOR UNRELATED REQUESTS Finally, some recipients seem to believe that offering support on one issue constitutes a standing offer to help with any and all future issues with a system. In the Google Analytics study, we were approached by one recipient that asked if we could help them solve a PHP error their Wordpress installation was showing, while other recipients asked for legal advice

20 Note that Twitter does not send mobile notifications for direct messages from people you do not follow or have not messaged with before. Instead they show up in a special section in the list of direct messages.

or recommendations for alternatives to Google Analytics. In these cases, a blanket refusal to help with anything unrelated to the study may be an appropriate policy, although it led to annoyance or disbelief from some recipients in our study.

8.4.3 *Positive and Negative Reactions*

There may also be very positive or negative reactions to the message.

The last two classes of messages a researcher may expect in a notification experiment are very positive or very negative reactions, which each have their own risks associated with them. Again, we briefly discuss our experiences.

Some recipients may offer gifts or payments as thanks, which can lead to ethical and legal problems.

GRATEFULNESS AND GIFTS In our studies, many recipients expressed their gratitude for our help. While this is generally positive, some also offered us gifts or payment. In these cases, researchers should be aware of the potential consequences of accepting gifts or payment for actions done as part of their job. In Germany, accepting gifts may constitute *Vorteilsnahme*, which can be reason for termination of the work contract and, in extreme cases, prosecution. We thus strongly recommend refusing all gifts that do not come in the form of formal donations to the university.

Such gifts may also be sent without prior warning.

In some cases, recipients did not even ask if they could send us a gift, and instead simply sent an unsolicited package containing anything from free postcards to branded mugs or bags and phone holders. Such unsolicited gifts pose an even greater challenge, as simply returning them is not always possible. In such cases, it may be advisable to discuss the matter with whoever is responsible for internal compliance and/or corruption prevention in the organization.

Other recipients may misunderstand the message, or threaten legal issue.

MISUNDERSTANDINGS AND THREATS The other side of the coin comes in the form of misunderstandings or legal threats. Some recipients interpreted the message as spam or scam messages, or even legal threats or defamatory, and threatened to send a cease-and-desist letter to university unless the messages stopped. In these cases, working with legal experts becomes even more important, as this can give certainty that the study does not break any laws. Nevertheless, the Google Analytics study led to an uncomfortable phone call with the chancellor of one involved university because of legal threats a website operator had made, and another set of calls with a company that had felt attacked by our messages and was complaining to our university, with which it had a cooperation agreement. Researchers should be prepared for such situations and be ready to defuse misunderstandings and remove participants from future notifications.

8.5 REMINDERS AND FOLLOW-UP

We discuss reminder and follow-up messages.

Several studies have used reminders to increase remediation rates, and our Google Analytics study showed that they can be effective. However, reminders and follow-up messages need to be used with care, as they may quickly exhaust the patience of recipients.

The first reminder can lead to additional remediations.

THE FIRST FOLLOW-UP In our third study, the first reminder led to significant improvements in the overall remediation. Some people reported that they either did not receive or did not trust the initial message, and only acted upon the reminder. On the other hand, other recipients became annoyed with the reminder and complained that they did not know why we kept contacting them. Overall, the reminder seems to have had a more positive than negative effect, so we would advise to consider sending a single reminder message.

Recipients become impatient after the second message, and may stop reading messages in detail.

THE SECOND FOLLOW-UP The second follow-up we sent was the final debriefing message that contained the information that the messages were part of a scientific study, and contained a link to a survey. Surprisingly, this message resulted in several complaints from recipients that told us that they had already remediated and that we should stop contacting them. This indicates that several recipients had stopped reading the message in detail, and started recognizing it based on the design and subject line, an effect that parallels the concept of *habituation* in psychology [88]. Stock *et al.* also reported that while their first reminder message was effective, the second no longer had any measurable effects on remediation [98], rendering further reminders pointless. We thus advise against sending more than three messages in total, as we would expect the negative reactions to increase with each message. We also recommend explicitly stating that the message is the final one the recipient will receive.

8.6 CONCLUSION

We hope this incomplete list of recommendations can be helpful for others.

This concludes our recommendations and lessons learned for future notification studies. Of course, as with any set of recommendations, the list is incomplete. However, we believe that explicitly formulating this knowledge can help to put future notification studies on a more solid basis, and avoid others having to re-learn the same lessons again and again. For further reading in this topic area, we once again refer to the best practice papers by Bajpai *et al.* [6], Cui *et al.* [25] and Durumeric *et al.* [33] for technical and methodological considerations. Vitak *et al.* [111] discusses the ethics of large-scale scans, as does the

companion paper to this chapter [72], and Marcia Hoffmann wrote about the legal aspects for Rapid7[21].

21 See https://www.rapid7.com/blog/post/2013/10/30/legal-considerations-for-widespread-scanning/, last accessed 2021-05-17.

9 CONCLUSION

In this dissertation, we have presented our work in detecting and driving remediation of security and privacy issues in multiple internet ecosystems. We first presented two automated public transparency systems for websites and email newsletters, respectively, and then used the data and technology of these systems to conduct three notification studies to evaluate different methods of promoting change.

We presented two transparency systems and three notification studies.

In Part i of the dissertation, we first gave an interdisciplinary overview of the challenges of ensuring the privacy and security of systems and whole ecosystems, considering the perspective of outsiders (i.e., scanning systems and notification studies) and insiders (i.e., the perspective of system administrators and the economic incentives of organizations). This knowledge informed the rest of the dissertation and allowed us to approach our research with an interdisciplinary perspective.

We gave an interdisciplinary overview over the field.

In Part ii, we presented two transparency tools, PrivacyScore and PrivacyMail, that perform automated privacy and security analyses in the web and email ecosystem. These tools, which are available online[1], have been received with interest by the users, sometimes observing over 500 visitors per day. They have also contributed to making the internet safer — data collected by PrivacyScore led to the detection and remediation of a vulnerability in a popular online shop system used by hundreds of pharmacies (cf. Section 3.5). PrivacyScore and PrivacyMail have served as the basis for several student theses and lab projects in our research group as well as at least four papers [54, 82, 93, 97] and one thesis written at other universities. They are also suitable for answering further research questions from a wide variety of different areas, from quantifying the presence of specific tracking technologies to performing user studies.

We presented two transparency tools with healthy user bases.

The technology of PrivacyScore also served as the basis for the three notification studies described in Part iii of the dissertation. Their common goal was to determine which methods can be used to incentivize system operators to make changes to their websites, investigating several factors like the message medium, the sender and framing of the message, and ranking the website's privacy- and security aspects in comparison to their peers and competitors. We found that the presence of such competitive rankings changes the responses of the recipients, but did not lead to large differences in how the websites were changed. This effect may have been an artifact of the used platform, as other

We used the developed technology for three notification studies investigating how to motivate system operators to improve their systems.

1 See https://privacyscore.org/ and https://privacymail.info, last accessed 2021-01-18.

studies have successfully leveraged competition in other areas [47, 104, 105, 118]. Changing the framing of the message achieved mixed results: more detailed explanations of the dangers of an information leak increased remediation for some issues, but decreased it for others, whereas framing an issue as one of compliance rather than a privacy issue significantly increased remediation rates. Contrary to previous studies, we also observed the sender of the message having an impact on remediation rates, and saw that reminder messages likely led to increased remediation rates. The most effective single change is a switch from email messages to letters, however, such a change may be impractical in some environments, as it requires manual address data collection and investment in postage.

We highlight the importance of leverage in driving change.

Based on our results, in combination with the previous studies in the field, we believe that one major factor in whether website operators remediate any misconfigurations is their fear of the consequences of inaction (i.e., the sender has *leverage* in the terminology of the transparency literature [53]). This can come from several sources (risk of compromise, legal requirements, loss of customers, etc.), and it may also lead to undesired reactions (e.g., threatening to sue if their website is not removed from a public ranking) if such actions are perceived to be the "cheaper" solution to the problem. Notification messages should also endeavor to avoid causing *too much* worry, as notification recipients may overreact (e.g., shut down their entire website instead of making a minor change) or experience significant psychological stress. Future research should take these results into consideration when planning new studies.

More studies are needed to understand effective notifications, and we support these studies with a collection of best practices.

We believe that the field of security and privacy notifications still has many open questions left — questions that are easier to answer when considered from an interdisciplinary lense. In addition to the existing overview of relevant aspects from other fields given in Chapter 2, we also included a dedicated chapter in Part iv of this dissertation, in which we share our experiences with preparing and conducting such studies. We highlight best practices and potential pitfalls in the different steps of conducting a notification study, ranging from the collection of the initial datasets and the preparation of the notification messages to more specific topics like the use of support tools, interactions with the recipients, and how to follow up if the first messages were unsuccessful. We hope that sharing our experience can allow future studies to start from a more solid foundation and avoid having to re-learn all the hard-won organizational and technical lessons of our studies.

Part V

APPENDIX

A COMPETITION

This appendix contains supplementary material for the competition study described in Chapter 5.

A.1 EXAMPLE NOTIFICATION MESSAGE

Below, we provide the text of the solicitation messages sent to the health insurance companies. As the messages for the two groups only differ in one sentence, we only show one version and highlight the differences with bold print in the text. We first provide the German version and then an English translation.

A.1.1 *German Version*

> [Anrede],
> mein Name ist Nicolas Walter, ich bin Master-Student der Technischen Universität Darmstadt mit Fachrichtung Wirtschaftsinformatik. Zurzeit schreibe ich meine Masterarbeit und führe in diesem Rahmen eine Studie zu Privatsphäre-Einstellungen auf Internetpräsenzen von Krankenkassen und Krankenversicherungen durch. Hierzu würde ich Ihnen gerne ein Universitätsprojekt vorstellen:
> Durch Zusammenarbeit von 6 Universitäten ist das Online-Tool PrivacyScore.org (https://privacyscore.org/) entstanden, mit dem es möglich ist Websites auf verschiedene Privatsphäre-Schwachstellen zu untersuchen. Wir untersuchen beispielsweise, wie gut Mail- und Web-Server verschlüsselt sind, inwieweit Tracking eingesetzt wird oder ob Schutz gegen verschiedene Angriffe besteht. Der Benefit für Unternehmen liegt darin, dass diese Informationen dazu genutzt werden können, die Privatsphäre für Kunden proakiv zu verbessern.
> Wir haben auch Ihre Unternehmenswebsite getestet. **Im Vergleich zu Ihren Wettbewerbern befinden Sie sich auf Platz [X] von 154, was die Privatsphäre-Einstellungen anbetrifft.** Der nachfolgende Screenshot zeigt ein paar Scanresultate **und Ihr Ranking**: [Screenshot]
> Alle Scans und Endergebnisse zu Ihrer Unternehmenswebsite können Sie auch selbst einsehen unter dem Link: [LINK]
> **[SCREENSHOT RANKING]**

Das Gesamtranking finden Sie unter folgendem Link: [LINK]

Für meine Studie, wäre ich Ihnen dankbar, wenn Sie mir zu den folgenden Fragen eine Rückmeldung geben könnten: Wie ist Ihre Haltung (aus Unternehmenssicht) gegenüber dieser Bewertung? Würden Sie Änderungen an Ihrer Webseite hin zu einer Privatsphäre-freundlicheren Version in Betracht ziehen?

Ich freue mich auf Ihre Rückmeldung und verbleibe mit freundlichen Grüßen,

Nicolas Walter

PS: Ich würde ich Sie bitten diese E-Mail an eine zuständige Abteilung weiterzuleiten, passend wäre möglicherweise die IT-Sicherheit oder der Datenschutz.

A.1.2 *English Version*

Dear [...],

my name is Nicolas Walter and I am a Master's student of the Technische Universität Darmstadt with a major in Business Informatics. I am currently conductng a study for my master thesis regarding privacy settings on the websites of health insurance companies and health insurers and I would like to introduce a university project.

In cooperation with 6 universities, the online tool Privacy-Score.org (https://privacyscore.org/) has been developed, which enables to examine websites for various privacy vulnerabilities. We examine, for example, how well mail and web servers are encrypted, to what extent tracking is used, or whether there is protec on against various attacks. Companies can benefit as they can use this information to proactively improve privacy for customers.

We have also tested your company website. **Compared to your competitors, you are ranked [x] out of 154 in terms of privacy settings.** The following screenshots show some scan results **and your ranking**: [SCREENSHOT RESULTS]

All scans and final results of your company's website can also be viewed here: [LINK]

[SCREENSHOT RANKING]

The overall ranking can be found under the following link: [LINK]

For my study, I would be glad if you could answer the following questions: what do you think about such an assessment from a company's point of view? And would

you consider to adapt your website in order to enhance the privacy settings?

I am looking forward to your feedback and remain with kind regards

Nicolas Walter

PS: Could you please forward this e-mail to a responsible department; the IT security or data protection department might be appropriate.

B

SECURITY NOTIFICATIONS

This appendix contains supplementary material for the security notification study described in Chapter 6.

B.1 EXAMPLE NOTIFICATION MESSAGES

Figure 16 shows an anonymized example letter in the original German version. In the following, we provide translated versions of the explanations of the different vulnerability types that were used in the letter. All messages started with the following text:

> To Whom it may concern,
> we are contacting you because you are listed as the responsible person for the website in its imprint. Within the context of a research project we found multiple vulnerabilities on your website http://www.example.com about which we would like to kindly advise you. The vulnerabilities stem from files that are publicly accessible, either unintentionally or by carelessness, and reveal sensible information. The following addresses are affected:

It was followed by a list of URLs and the relevant explanations, listed below. The message then closed by stating:

> In the interest of your websites security we advise you to remedy those vulnerabilities as soon as possible. Further information on our project, your vulnerabilities, assistance in remediation and a status check for your website is available at: [URL with token]
> I will be happy to assist you with any further questions.
> With kind regards,
> Max Maass

The message signature contained the name, fax number, email and postal address of the sender. We now provide the individual text blocks that describe the different vulnerabilities.

B.1.1 *SSH Key*

BASELINE At this address anyone can download an access key which can presumably be used to log in to your website and get full access. Please consult an expert for the next steps, because we can not determine the full impact of this problem.

TECHNISCHE UNIVERSITÄT DARMSTADT

Max Maass | SEEMOO / TU Darmstadt | Mornewegstr. 32 | 64293 Darmstadt

[redacted]

Schwachstellen auf Ihrer Website [redacted]

Sehr geehrte Damen und Herren,

wir wenden uns an Sie, da Sie im Impressum der Webseite als Verantworliche genannt sind. Im Rahmen einer Forschungsarbeit haben wir mehrere Schwachstellen auf Ihrer Website [redacted] gefunden, über die wir Sie gerne in Kenntnis setzen möchten. Bei den Schwachstellen handelt es sich um Dateien, die unbeabsichtigt oder fahrlässigerweise öffentlich zugänglich sind und sensible Informationen preisgeben.
Es handelt sich um folgende Adressen:

[redacted]`/dump.sql`
Über diese Adresse lässt sich eine Sicherung Ihrer Datenbank herunterladen. Obwohl wir deren Inhalte nicht im Einzelnen überprüft haben, lässt sich davon ausgehen, dass darin nicht für die Öffentlichkeit bestimmte Inhalte enthalten sind.

[redacted]`/info.php`
Über diese Adresse lassen sich Informationen über verwendete Software und deren Versionen, als auch interne Konfigurationsdetails Ihres Servers abrufen, die aus Sicherheitsgründen nicht öffentlich verfügbar sein sollten.

Im Sinne der Sicherheit Ihrer Webseite raten wir Ihnen, diese Schwachstellen schnellstmöglich zu beheben bzw. Selbiges zu veranlassen.
Weitergehende Informationen zum Projekt, den Schwachstellen, Hilfestellungen zur Behebung und eine Statusabfrage für Ihre Webseite finden Sie unter:

`https://web-survey.seemoo.tu-darmstadt.de/`[redacted]

Für Fragen stehe ich sehr gerne zur Verfügung.

Mit freundlichen Grüßen

[redacted]
Max Maass

Fachbereich Informatik
Sichere Mobile Netze
Secure Mobile Networking

SEEMOO SECURE MOBILE NETWORKING

Max Maass

SEEMOO / TU Darmstadt
Mornewegstraße 32
64293 Darmstadt

Fax +49 6151 16 - 25471
web-survey@seemoo.tu-darmstadt.de

Datum: 11.6.2018

CRISP Center for Research in Security and Privacy

Privacy and Trust for Mobile Users

Figure 16: German example notification letter, anonymized for release (translation provided below).

ATTACK This vulnerability was only observed once and thus does not contain a version with attack scenario.

B.1.2 *TLS Key*

BASELINE At this address anyone can access the private key to your websites transport encryption. Because of that we have to assume that the encryption can actually be abrogated. Please deactivate the access to the key, renew the key as well as your encryption certificate and revoke the old key.

ATTACK This vulnerability was only observed once and thus does not contain a version with attack scenario.

B.1.3 *Database Backup*

BASELINE At this address, anyone can download a backup of your database. Although we did not review its content in detail, it is very likely that its contents are not intended for public consumption.

ATTACK An attacker can likely extract the list of users from that backup and can possibly extract the passwords. This may grant them full access to the website and allow them to manipulate the content, which can lead to defamation.

B.1.4 *VCS*

BASELINE The availability of this file indicates that the source code of your website is publicly accessible. While we did not verify those contents in detail, we assume that it contains content that is not meant to be publicly accessible, such as login and contact data or internal configuration files.

ATTACK Depending on whether an attacker discovers information like login or contact data, she can in some circumstances acquire full access to your website or impersonate you with the help of the contact data to gain access to further information by fraud.

B.1.5 *Server-Status*

BASELINE At this address anyone can see which pages on the website are currently accessed from which IP address using which parameters. This means that in doing so you illicitly disclose the identity and activities of your visitors.

Table 14: Requested paths for the different vulnerabilities.

Vulnerability	Paths
Keyfile	`id_rsa, .ssh/id_rsa, privatekey.key, private.key, myserver.key, key.pem, privkey.pem, [domain].key, [domain_full].key, [subdomain].key, [domain].pem, [full_domain].pem, [subdomain].pem, cert.pem, certificate.pem, domain.key`
Database	`dump.db, dump.sql, sqldump.sql, sqldump.db, db.sqlite, data.sqlite, sqlite.db, [domain].sql, [domain_full].sql, [subdomain].sql, [domain].db, [domain_full].db, [subdomain].db`
Core dump	`core`
VCS	`.git/HEAD, .svn/wc.db`
Status	`server-status/, server-info/`
PHPInfo	`phpinfo.php, test.php, info.php`

ATTACK Using this information an attacker can trace who visits your website and learn about visit duration and links clicked. In the worst case she thereby learns the so called "Session ID", which enables her to seize the role of a visitor and impersonate them.

B.1.6 *Server-Info*

BASELINE At this address anyone can retrieve information about software modules in use as well as their versions and internal configuration details of your server which should not be public for security reasons.

ATTACK With the help of such version information an attacker can very easily determine whether outdated software with known vulnerabilities is in use. If this is the case, she can exploit those easily and in the worst case gain access to the server.

B.1.7 *PHPInfo*

BASELINE At this address anyone can access information about the software in use as well as internal configuration details, which should not be publicly accessible for security reasons.

ATTACK With the help of such version information an attacker can easily determine whether outdated software with known vulnerabilities is in use. If this is the case, she can exploit them with ease and in the worst case gain access to the server.

C

COMPLIANCE

This appendix contains supplementary material for the notification study described in Chapter 7.

C.1 GOOGLE ANALYTICS MISCONFIGURATION

Activating IP anonymization for Google Analytics is a non-trivial process. It has to be activated explicitly and by the website operator through a change in the source code of the website. The exact required changes depend on the used Google Analytics library, as the system can be embedded in multiple different ways. Listing 1 shows a number of potential mistakes that can be made when activating IP anonymization using the analytics.js library. The setting string is case-sensitive and named inconsistently between the different versions of the library, and using an incorrect string will not raise any errors (attempt 2 in the listing). Finally, the setting needs to be activated at the right time in the program flow: after the tracking ID was set (attempt 1), but before the tracking requests are sent (attempt 3). Thus, despite trying three times to activate IP anonymization, the code from Listing 1 would result in a website operating without anonymization. Such errors led to significant confusion from the notification recipients, which required detailed explanations to remedy. Part of the remediation instructions given by the self-service are shown in Figure 17.

Listing 1: Examples of erroneous IP anonymization configurations for Google Analytics using analytics.js

```
1  ga('set', 'anonymizeIp', true);  // Attempt 1
2  // Error: must be done after configuring the tracking ID
3  ga('create', 'UA-XXXXX-Y', 'auto');
4  // Configure the tracking ID
5  ga('set', 'anonymizeIP', true);  // Attempt 2
6  // Error: must be spelled 'anonymizeIp'
7  ga('send', 'pageview');
8  // Send the pageview
9  ga('set', 'anonymizeIp', true);  // Attempt 3
10 // Error: must be done before sending the pageview
```

Aktivieren der IP-Anonymisierung

Die Methode zur Aktivierung der IP-Anonymisierung unterscheidet sich je nach verwendeter Google Analytics-Bibliothek. Im Folgenden haben wir die korrekten Methoden für die häufigsten Bibliotheken zusammengetragen.

Bei der Nutzung der analytics.js-Bibliothek kann die IP-Anonymisierung mit der „anonymizeIp"-Option aktiviert werden (Zeile 7):

```
1. (function(i,s,o,g,r,a,m){i['GoogleAnalyticsObject']=r;i[r]=i[r]||function(){
2. (i[r].q=i[r].q||[]).push(arguments)},i[r].l=1*new Date();a=s.createElement(o),
3. m=s.getElementsByTagName(o)[0];a.async=1;a.src=g;m.parentNode.insertBefore(a,m)
4. })(window,document,'script','https://www.google-analytics.com/analytics.js','ga');
5.
6. ga('create', 'UA-XXXXX-Y', 'auto');
7. ga('set', 'anonymizeIp', true);  // Anonymisierung aktivieren. Muss vor
8.                                  // ga('send', 'pageview') platziert werden.
9. ga('send', 'pageview');
```

Wenn die Webseite noch die veraltete ga.js-Bibliothek verwendet, wird die Anonymisierung mit der _gat._anonymizeIp-Funktion aktiviert (Zeile 3):

```
1. var _gaq = _gaq || [];
2. _gaq.push(['_setAccount', 'UA-XXXXX-X']);
3. _gaq.push(['_gat._anonymizeIp']);  // Anonymisierung aktivieren. Muss vor
4.                                    // _gaq.push(['_trackPageview']) platziert werden.
5. _gaq.push(['_trackPageview']);
6.
7. (function() {
8.   var ga = document.createElement('script'); ga.type = 'text/javascript'; ga.async = true;
9.   ga.src = ('https:' == document.location.protocol ? 'https://ssl' : 'http://www') + '.google-analytics.com/ga.js';
10.  var s = document.getElementsByTagName('script')[0]; s.parentNode.insertBefore(ga, s);
11. })();
```

Wenn die Webseite die gtag.js-Bibliothek verwendet, kann die IP-Anonymisierung mit einem weiteren Parameter in der Konfiguration aktiviert werden (Zeile 7):

```
1. <script async src="https://www.googletagmanager.com/gtag/js?id=GA_MEASUREMENT_ID"></script>
2. <script>
3.   window.dataLayer = window.dataLayer || [];
4.   function gtag(){dataLayer.push(arguments);}
5.   gtag('js', new Date());
6.
7.   gtag('config', 'GA_MEASUREMENT_ID', {'anonymize_ip': true});  // Der weitere Parameter {'anonymize_ip': true}
8.                                                                 // aktiviert die Anonymisierung
9. </script>
```

Wenn Sie den Google Tag Manager verwenden, um Ihre Tracking-Skripte zu verwalten, aktivieren Sie die IP-Adressen-Anonymisierung, indem Sie das entsprechende Tag der Google-Tag-Manager-Oberfläche editieren. Dort können Sie das Feld „anonymizeIp" hinzufügen und auf den Wert „true" setzen.

Weitere Informationen zum Effekt und der technischen Umsetzung erhalten Sie direkt von Google.

Figure 17: Part of the instructions on activating IP Anonymization, as given in the self-service tool.

C.2 BEHAVIOR OF CO-OWNED WEBSITES

In the methodology section, we note that we assume that a notified operator in charge of more than one domain will likely remediate either all or none of their websites. We thus decided to introduce a weighing

mechanism to reduce the impact of such correlated remediations on the overall results (cf. Section 7.2.7).

Here, we briefly attempt to investigate the veracity of this assumption. In total, our dataset contained 88 recipients that controlled more than one website. Of these, 77 (87.5 %) had indeed remediated either all (40) or none (37) of their websites at the end of the study timeframe (this count includes operators from the control group). Of the 40 fully compliant recipients, 34 (85 %) needed a timespan of less than 2 days between the first and last remediation. These results indicate that while not every website operator fulfilled the assumption of correlated remediations, a large majority did, and the impact of the remaining operators on the overall results should be small.

C.3 EXAMPLE NOTIFICATION MESSAGES

This section contains a translated version of the different messages that were sent to recipients. It is quoted verbatim from the online supplementary material for the paper [73], with minor adjustments to formatting. An example letter from the UNI-CS – GDPR+FINE group is shown in Figure 18.

University Groups

Dear Madams and Sirs,

we are contacting you because you are listed in the imprint of the following website[s] as the responsible party: [single URL or list of URLs]

As part of a research project, we have recently inspected your website, and found that you are using Google Analytics without anonymizing the IP addresses of your visitors.

PRIVACY ONLY Due to the lack of anonymization, you are violating the privacy of your visitors, as the full IP address is saved by Google. This allows Google to collect more data than absolutely necessary. The full IP addresses can be used by Google to create more precise profiles of your visitors.

To improve the privacy of your users, we recommend activating the IP Anonymization feature, or to use your own web analytics system, like the free web analytics software "Matomo".

GDPR AND GDPR+FINE ONLY Operating the website[s] with this configuration is not compliant with the law.

There is a violation of Art. 6 Para. 1 lit. f GDPR, as the interests of the website visitors predominate when weighing up the interests. You have not pseudonymised or anonymised the IP address of the website visitors. This configuration of the Google Analytics tool violates the

TECHNISCHE UNIVERSITÄT DARMSTADT

Max Maass | Secure Mobile Networking Lab | Mornewegstr. 32 | 64293 Darmstadt

[redacted]

Betrifft: Ihre Webseite [redacted]

Sehr geehrte Damen und Herren,

Wir wenden uns an Sie, da Sie im Impressum der o.g. Webseite als Verantwortliche genannt sind. Im Rahmen einer Forschungsarbeit haben wir vor einigen Tagen die im Betreff genannte Webseite untersucht. Dabei haben wir festgestellt, dass Sie Google Analytics einsetzen. Allerdings verwenden Sie die von Google Analytics angebotene Funktion zur Anonymisierung der IP-Adressen nicht korrekt, weswegen diese nicht die gewünschte Wirkung entfaltet.

Der Betrieb der Webseite ist in ihrer jetzigen Konfiguration nach vorherrschender Meinung rechtlich unzulässig.

Es liegt ein Verstoß gegen Art. 6 Abs. 1 lit. f DSGVO vor, da im Rahmen einer Interessenabwägung die Interessen der Webseitenbesucher überwiegen. Sie haben keine Pseudonymisierung bzw. Anonymisierung der IP-Adresse der Webseitenbesucher vorgenommen. Diese Konfiguration des Google Analytics Tools verstößt gegen die Grundsätze der Datenminimierung, Speicherbegrenzung und der Nutzung von Pseudonymisierung bzw. Anonymisierung. Ein solcher Verstoß kann gemäß Art. 83 DSGVO mit einer Geldbuße geahndet werden.

Sanktionen sind in Art. 83 sowie Art. 84 DSGVO geregelt. Demnach können die Aufsichtsbehörden bei datenschutzrechtlichen Verstößen Geldbußen verhängen. Diese Geldbußen müssen in jedem Einzelfall wirksam, verhältnismäßig und abschreckend sein (vgl. Art. 83 Abs. 1 DSGVO). Bei Verstößen gegen die Grundsätze für die Verarbeitung gemäß Art. 5 DSGVO (Grundsätze für die Verarbeitung personenbezogener Daten) sowie Art. 6 DSGVO (Rechtmäßigkeit der Verarbeitung) können Geldbußen von bis zu 20 Mio. € oder im Fall eines Unternehmens von bis zu 4 % des gesamten weltweit erzielten Jahresumsatzes des vorangegangenen Geschäftsjahres verhängt werden, je nachdem, welcher der Beträge höher ist (Art. 83 Abs. 5 lit. a DSGVO).

Zur Prüfung Ihrer Google-Analytics-Konfiguration bietet die Universität Bamberg ein kostenloses Tool an, das Sie unter *https://checkgoogleanalytics.psi.uni-bamberg.de* erreichen können. Dort finden Sie auch weitere Informationen, wie Sie die Probleme beheben können.

Falls Sie Fragen haben, finden Sie unsere Kontaktdaten in der Seitenspalte. Wir möchten Sie bitten, sich nach Möglichkeit via E-Mail oder Brief zu melden, da das Telefon nur Dienstag Vormittag von 9–11 Uhr besetzt ist, und in Ihrem Schreiben die Webseite(n) zu nennen, auf die Sie sich beziehen.

Mit freundlichen Grüßen

Max Maass

Fachbereich Informatik
Sichere Mobile Netze
Secure Mobile Networking

SEEMOO SECURE MOBILE NETWORKING

Max Maass
SEEMOO / TU Darmstadt
Mornewegstraße 32
64293 Darmstadt

notification@seemoo.tu-darmstadt.de
Tel. +49 6151 16 - 25473

CRISP Center for Research in Security and Privacy

Privacy and Trust for Mobile Users

CROSSING

Figure 18: German example notification letter from the UNI-CS – GDPR+FINE group, anonymized for release (translation provided below).

principles of data minimization, storage limitation and the use of pseudonymization or anonymization.

GDPR+FINE ONLY Such an infringement may be punishable by a fine in accordance with Art. 83 DSGVO.

Sanctions are regulated in Art. 83 and Art. 84 DSGVO. According to this, the supervisory authorities can impose fines in the event of violations of data protection law. These fines must be effective, proportionate and dissuasive in each individual case (cf. Art. 83 (1) DSGVO). In the event of infringements of the principles governing processing pursuant to Art. 5 DSGVO (principles governing the processing of personal data) and Art. 6 DSGVO (lawfulness of processing), fines of up to € 20 million or, in the case of a company, up to 4 % of the total annual worldwide turnover achieved in the previous financial year, whichever is the higher, may be imposed (Art. 83 para. 5 lit. a DSGVO).

ALL FRAMINGS To check your Google Analytics configuration, the University of Bamberg provides the tool "Check Google Analytics", which is operated by the Privacy and Security in Information Systems group (https://www.uni-bamberg.de/psi/). This tool also contains information on how the issue can be remediated.

If you have any questions, you can find my contact information in the signature of this email. Be advised that the phone calls are only taken Thursday, from 9-11 am.

Sincerely, [Name of the sender]

[Email signature: name, affiliation, address, phone number]

Private Sender

Dear Madams and Sirs,

when visiting your website[s], I noticed that you are using Google Analytics without anonymizing the IP address of your visitors. This applies to the following site[s]: [single URL or list of URLs]

PRIVACY ONLY Due to the lack of anonymization, you are violating the privacy of your visitors, as the full IP address is saved by Google. This allows Google to collect more data than absolutely necessary. The full IP addresses can be used by Google to create more precise profiles of your visitors. This supports Google in its highly questionable tracking practices, which I absolutely reject.

To improve the privacy of your users it is recommended to activate IP Anonymisation or to use your own analysis system such as the free web analysis tool "Matomo".

GDPR AND GDPR+FINE [Identical to the university groups.]

ALL FRAMINGS To check your Google Analytics configuration, the University of Bamberg provides the tool "Check Google Analytics", which is operated by the Privacy and Security in Information Systems group (https://www.uni-bamberg.de/psi/). This tool also contains information on how the issue can be remediated.

I am happy to answer any questions you may have — simply reply to this email.

Sincerely, Max Maass

C.4 SURVEY

This section contains a translated version of the survey that was sent to notification recipients. It is quoted verbatim from the online supplementary material of the paper [73].

1. **Dear Participant,**
   Thank you for participating in this 5-minute survey. By doing so, you are supporting a research project of Technische Universität Darmstadt, Otto-Friedrich-Universität Bamberg, and Goethe Universität Frankfurt on data protection. Further information on data protection in this survey can be found here: [link to information on the study and on data protection]

   O I have read the notes to the study and agree to participate in the study. I agree that the data collected in the course in the study may be evaluated for scientific purposes and stored and published in anonymised form. I am aware that my participation is voluntary and that I can stop the trial at any time and without giving reasons.

   O I do **not** agree

2. **We sent you this notification at the beginning of July. Did you read this notification?**

   [Picture of a Notification]

3. **What was your first impression of our notification?**
   The notification makes a trustworthy impression.

O	O	O	O
Strongly disagree	Disagree	Agree	Strongly agree

   On the basis of the notification I could understand the problem of the missing IP Anonymization.

O	O	O	O
Strongly disagree	Disagree	Agree	Strongly agree

4. **Please briefly describe which aspects of the notification led you to trust it.**
   a) Aspect .....
   b) Aspect .....
   c) Aspect .....

5. **Please briefly describe which aspects of the notification led to having *NO/LESS* confidence in the notification.**
   a) Aspect .....
   b) Aspect .....
   c) Aspect .....

6. **Have you fixed the problem of missing IP anonymisation of Google Analytics on your website before 27.09.2019?**
   O Yes
   O No
   O I don't know
   O Page was completely shut down

7. **Since the problem of missing IP Anonymization has already been solved on your website: which of the following statements best applies to you?**
   O I fixed the problem myself without help.
   O I fixed the problem myself with help.
   O I have forwarded the problem/issue to colleagues in my organization
   O I have asked my external service provider to fix the problem.
   O I have hired a new service provider to fix the problem.
   O Other: .....

8. **Can you imagine what could be reasons why you did not activate the IP anonymisation of Google Analytics on your website until 27.09.2019?** *Multiple answer possible*
   ☐ Problem was not known
   ☐ Lack of knowledge on how to solve the problem
   ☐ Missing time
   ☐ Problem has no priority

- ☐ Notification did not seem serious to me
- ☐ Other: .....

9. **Did you know *BEFORE* we notified you that you were using Google Analytics on your website?**
   - O Yes
   - O No
10. **Had you heard about the IP Anonymization feature *BEFORE* we notified you?**
   - O Yes, and I knew the purpose
   - O Yes, I didn't know exactly what it was
   - O No
11. **Did you know *BEFORE* we notified you about the missing IP Anonymization of Google Analytics on your website?**
   - O If yes: what do you think were the reasons for the lack of IP Anonymization?
     .....
   - O No
12. **During our study (July to September 2019), various media reported a verdict of the regional court Dresden reports, which states that the activation of IP anonymisation at Google Analytics is necessary for legally compliant operation. Which of the statements apply to you?** *Multiple answer possible*
   - ☐ The judgement was decisive for activating IP Anonymization.
   - ☐ The IP Anonymization was activated INDEPENDENTLY of the judgement.
   - ☐ I haven't heard about the verdict yet.
   - ☐ Other: .....
13. **In our notification we referred you to the CheckGA tool of Otto-Friedrich-Universität Bamberg.**

    [Picture of the tool]

    How helpful did you find the CheckGA tool?

not helpful at all				extremely helpful	I have not used the tool	I do not know the tool
O	O	O	O	O	O	O

14. **Which of the following check tools are known to you or have you already used?** *Multiple answer possible*

	Known	Known + Used	Unknown
Qualys SSL Check	O	O	O
Webkoll	O	O	O
Mozilla Observatory	O	O	O
HTBridge	O	O	O
Immuniweb	O	O	O
Other:	O	O	O

15. **Would you like to receive future notifications about privacy issues on your website?**

    O Yes

    O No

16. **In which way would you like to be informed about data protection problems?** *Multiple answer possible*

    ☐ Email

    ☐ Letter

    ☐ Call

    ☐ Blogposts

    ☐ Other: .....

17. **Would you be willing to pay for such notifications?**

    O Yes

    O No

18. **Organizational data**

    How many employees does your company or organization have? .....

19. **Who in your organisation is responsible for maintaining the website?**

    O Employees in the organisation (please indicate how many employees are responsible for this) .....

    O Website support by an external agency

20. **Is there anything else you would like to tell us? For example, something that we should consider for future notifications.**

21. **Interest in research results/future study participation**
Note: Your e-mail address will be stored separately from all other information in this survey.

    ☐ Yes, I would like to be informed about the results of this study.

    ☐ Yes, I would like to be informed about further studies.

Many thanks for your participation! We would like to thank you very much for your assistance. Your answers have been saved, you can now close the browser window.

BIBLIOGRAPHY

[1] Gunes Acar, Steven Englehardt, and Arvind Narayanan. "No boundaries: data exfiltration by third parties embedded on web pages." In: *Proceedings on Privacy Enhancing Technologies* 2020.4 (2020), pp. 220–238.

[2] Gunes Acar, Christian Eubank, Steven Englehardt, Marc Juarez, Arvind Narayanan, and Claudia Diaz. "The Web Never Forgets: Persistent Tracking Mechanisms in the Wild." In: *Proceedings of the 2014 ACM SIGSAC Conference on Computer and Communications Security – CCS '14* (2014), pp. 674–689.

[3] Mike Ananny and Kate Crawford. "Seeing without knowing: Limitations of the transparency ideal and its application to algorithmic accountability." In: *New Media & Society* 20.3 (2016), pp. 973–989.

[4] Amelia Andersdotter and Anders Jensen-Urstad. "Evaluating Websites and Their Adherence to Data Protection Principles: Tools and Experiences." In: *IFIP Advances in Information and Communication Technology*. Vol. 498. 2016, pp. 39–51.

[5] Joshua D Angrist. "The Perils of Peer Effects." In: *Labour Economics* 30 (2014), pp. 98–108.

[6] Vaibhav Bajpai, Anna Brunstrom, Anja Feldmann, Wolfgang Kellerer, Aiko Pras, Henning Schulzrinne, Georgios Smaragdakis, Matthias Wählisch, and Klaus Wehrle. "The Dagstuhl beginners guide to reproducibility for experimental networking research." In: *SIGCOMM Computer Communication Review* 49.1 (2019).

[7] Johannes M Bauer and Michel J G van Eeten. "Cybersecurity: Stakeholder incentives, externalities, and policy options." In: *Telecommunications Policy* 33.10 (2009), pp. 706–719.

[8] France Bélanger and Robert E Crossler. "Privacy in the Digital Age: A Review of Information Privacy Research in Information Systems." In: *The Mississippi Quarterly* 35.4 (2011), pp. 1017–1041.

[9] Fabien C Y Benureau and Nicolas P Rougier. "Re-run, Repeat, Reproduce, Reuse, Replicate: Transforming Code into Scientific Contributions." In: *Frontiers in Neuroinformatics* 11 (2017), p. 69.

[10] Rainer Böhme. "Security Metrics and Security Investment Models." In: *Advances in Information and Computer Security*. 2010, pp. 10–24.

[11] Sven Braun and Anne-Marie Oostveen. "Encryption for the masses? An analysis of PGP key usage." In: *Mediatization Studies* 2.0 (2019), pp. 69–84.

[12] Matthias Brecht and Thomas Nowey. "A Closer Look at Information Security Costs." In: *The Economics of Information Security and Privacy*. Ed. by Rainer Böhme. 2013, pp. 3–24.

[13] Norman Breslow. "Analysis of Survival Data under the Proportional Hazards Model." In: *International statistical review* 43.1 (1975), pp. 45–57.

[14] Davide Canali, Davide Balzarotti, and Aurélien Francillon. "The role of web hosting providers in detecting compromised websites." In: *Proceedings of the 22nd International Conference on World Wide Web – WWW '13*. 2013, pp. 177–187.

[15] Orçun Çetin, Lisette Altena, Carlos Gañán, and Michel Van Eeten. "Let Me Out! Evaluating the Effectiveness of Quarantining Compromised Users in Walled Gardens." In: *Proceedings of the 14th Symposium on Usable Privacy and Security – SOUPS '18*. 2018, pp. 251–263.

[16] Orçun Çetin, Carlos Ganan, Lisette Altena, Samaneh Tajalizadehkhoob, and Michel van Eeten. "Tell Me You Fixed It: Evaluating Vulnerability Notifications via Quarantine Networks." In: *IEEE European Symposium on Security and Privacy 2019 – EuroS&P '19*. 2019, pp. 326–339.

[17] Orçun Çetin, Carlos Ganan, Maciej Korczynski, and Michel van Eeten. "Make Notifications Great Again: Learning How to Notify in the Age of Large-Scale Vulnerability Scanning." In: *Proceedings of the 16th Annual Workshop on the Economics of Information Security – WEIS '17*. 2017, p. 23.

[18] Orçun Çetin, Mohammad Hanif Jhaveri, Carlos Gañán, Michel van Eeten, and Tyler Moore. "Understanding the role of sender reputation in abuse reporting and cleanup." In: *Journal of Cybersecurity* 2.1 (2016), pp. 83–98.

[19] Lars Thøger Christensen and George Cheney. "Peering into transparency: Challenging ideals, proxies, and organizational practices." In: *Communication theory: CT: a journal of the International Communication Association* 25.1 (2015), pp. 70–90.

[20] Wolfie Christl. *Corporate Surveillance in Everyday Life*. Tech. rep. Cracked Labs, 2017. URL: https://crackedlabs.org/en/corporate-surveillance.

[21] Jiska Classen, Daniel Wegemer, Paul Patras, Tom Spink, and Matthias Hollick. "Anatomy of a vulnerable fitness tracking system: Dissecting the fitbit cloud, app, and firmware." In: *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 2.1 (2018).

[22] *Court of Justice of the EU, Judgment ECLI:EU:C:2018:388*. URL: http://curia.europa.eu/juris/liste.jsf?num=C-210/16 (visited on 06/09/2020).

[23] David R Cox. "Regression Models and Life-Tables." In: *Journal of the Royal Statistical Society. Series B, Statistical methodology* 34.2 (1972), pp. 187–220.

[24] Dave Crocker. *Mailbox Names for Common Services, Roles and Functions*. RFC 2142. RFC Editor, 1997. URL: https://www.rfc-editor.org/rfc/rfc2142.txt.

[25] Ang Cui and Salvatore J Stolfo. "Reflections on the engineering and operation of a large-scale embedded device vulnerability scanner." In: *Proceedings of the First Workshop on Building Analysis Datasets and Gathering Experience Returns for Security - BADGERS '11*. 2011. DOI: 10.1145/1978672.1978674.

[26] Cameron Davidson-Pilon et al. *Lifelines v0.25.4*. 2020. URL: https://doi.org/10.5281/zenodo.4002777.

[27] Martin Degeling, Christine Utz, Christopher Lentzsch, Henry Hosseini, Florian Schaub, and Thorsten Holz. "We Value Your Privacy ... Now Take Some Cookies: Measuring the GDPR's Impact on Web Privacy." In: *Proceedings of the 2019 Network and Distributed System Security Symposium – NDSS '19* (2019).

[28] Tim Dierks and Eric Rescorla. *The Transport Layer Security (TLS) Protocol Version 1.2*. RFC 5246. RFC Editor, 2008. URL: https://www.rfc-editor.org/rfc/rfc5246.txt.

[29] Serigne Mouhamadane Diop, Jema David Ndibwile, Doudou Fall, Shigeru Kashihara, and Youki Kadobayashi. "To Coerce or Not to Coerce? A Quantitative Investigation on Cybersecurity and Cybercrime Legislations Towards Large-Scale Vulnerability Notifications." In: *International Conference on Software Reliability Engineering Workshops – ISSRE '19 Workshops*. 2019.

[30] Wim Dubbink, Johan Graafland, and Luc Van Liedekerke. "CSR, Transparency and the role of intermediate organisations." In: *Journal of Business Ethics* 82.2 (2008), pp. 391–406.

[31] Zakir Durumeric, David Adrian, Ariana Mirian, Michael Bailey, and J Alex Halderman. "A Search Engine Backed by Internet-Wide Scanning." In: *Proceedings of the 22nd ACM SIGSAC Conference on Computer and Communications Security – CCS '15*. 2015.

[32] Zakir Durumeric, David Adrian, Ariana Mirian, James Kasten, Elie Bursztein, Nicolas Lidzborski, Kurt Thomas, Vijay Eranti, Michael Bailey, and J Alex Halderman. "Neither Snow Nor Rain Nor MITM...: An Empirical Analysis of Email Delivery Security." In: *Proceedings of the 2015 Internet Measurement Conference – IMC '15*. 2015, pp. 27–39.

[33] Zakir Durumeric, Eric Wustrow, and J Alex Halderman. "ZMap: Fast Internet-wide scanning and its security applications." In: *22nd USENIX Security Symposium – USENIX Security '13*. 2013.

[34] Zakir Durumeric et al. "The Matter of Heartbleed." In: *Proceedings of the 2014 Internet Measurement Conference – IMC '14*. 2014, pp. 475–488.

[35] Brad Efron. "Bootstrap Methods: Another Look at the Jackknife." In: *The Annals of Statistics* 7.1 (1979), pp. 1–26.

[36] Steven Englehardt. "Automated Discovery of Privacy Violations on the Web." PhD thesis. Princeton University, 2018.

[37] Steven Englehardt, Jeffrey Han, and Arvind Narayanan. "I never signed up for this! Privacy implications of email tracking." In: *Proceedings on Privacy Enhancing Technologies* 2018.1 (2018), pp. 109–126.

[38] Steven Englehardt and Arvind Narayanan. "Online Tracking: A 1-million-site Measurement and Analysis." In: *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security – CCS '16*. 2016, pp. 1388–1401.

[39] Steven Englehardt, Dillon Reisman, Christian Eubank, Peter Zimmerman, Jonathan Mayer, Arvind Narayanan, and Edward W Felten. "Cookies That Give You Away." In: *Proceedings of the 24th International Conference on World Wide Web – WWW '15* (2015), pp. 289–299.

[40] Ernst Fehr and Simon Gächter. "Altruistic punishment in humans." In: *Nature* 415 (2002).

[41] Leon Feistinger. "A Theory of Social Comparison Processes." In: *Human Relations* 7.2 (1954).

[42] Nathaniel Fruchter, Hsin Miao, Scott Stevenson, and Rebecca Balebako. "Variations in Tracking in Relation to Geographic Location." In: *Web 2.0 Security & Privacy (W2SP)* (2015).

[43] Jin P Gerlach, Nicole Eling, Nora Wessels, and Peter Buxmann. "Flamingos on a slackline: Companies' challenges of balancing the competing demands of handling customer information and privacy." In: *Information Systems Journal* 29.2 (2018), pp. 1–28.

[44] Kathleen E Greenaway and Yolande E Chan. "Theoretical Explanations for Firms' Information Privacy Behaviors." In: *Journal of the Association for Information Systems* 6.6 (2005), pp. 171–198.

[45] Kathleen E Greenaway, Yolande E Chan, and Robert E Crossler. "Company information privacy orientation: a conceptual framework." In: *Information Systems Journal* 25.6 (2015), pp. 579–606.

[46] Johannes Haupt, Benedict Bender, Benjamin Fabian, and Stefan Lessmann. "Robust identification of email tracking: A machine learning approach." In: *European Journal of Operational Research* 271.1 (2018), pp. 341–356.

[47] Shu He, Gene Moo Lee, Sukjin Han, and Andrew B Whinston. "How would information disclosure influence organizations' outbound spam volume? Evidence from a field experiment." In: *Journal of Cybersecurity* 2.1 (2016), pp. 99–118.

[48] Sture Holm. "A Simple Sequentially Rejective Multiple Test Procedure." In: *Scandinavian Journal of Statistics* 6.2 (1979), pp. 65–70.

[49] Ralph Holz, Johanna Amann, Olivier Mehani, Matthias Wachs, and Mohamed Ali Kaafar. "TLS in the wild: An internet-wide analysis of TLS-based protocols for electronic communication." In: *Proceedings of the 2016 Network and Distributed System Security Symposium – NDSS '16*. 2016.

[50] Hang Hu, Peng Peng, and Gang Wang. "Characterizing Pixel Tracking through the Lens of Disposable Email Services." In: *Proceedings of the 2019 IEEE Symposium on Security & Privacy – S&P '19*. 2019, pp. 545–559.

[51] Nicola Jentzsch. "State-of-the-Art of the Economics of Cyber-Security and Privacy." In: *Innovation Framework for Privacy and Cyber-security Market Opportunities – Deliverable D*. 2016.

[52] Haojian Jin, Minyi Liu, Kevan Dodhia, Yuanchun Li, Gaurav Srivastava, Matthew Fredrikson, Yuvraj Agarwal, and Jason I Hong. "Why Are They Collecting My Data? Inferring the Purposes of Network Traffic in Mobile Apps." In: *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 2.4 (2018), pp. 1–27.

[53] Sara Kahn-Nisser. "Constructive Criticism: Shaming, Incentives, and Human Rights Reforms." In: *Politics and Policy* 46.1 (2018), pp. 58–83.

[54] Shirin Kalantari, Andreas Put, and Bart De Decker. "Trackers in Your Inbox: Criticizing Current Email Tracking Practices." In: *Privacy Technologies and Policy*. 2021.

[55] Georgios Kambourakis, Gerard Draper Gil, and Ignacio Sanchez. "What Email Servers Can Tell to Johnny: An Empirical Study of Provider-to-Provider Email Security." In: *IEEE Access* 8 (2020).

[56] Edward Kaplan and Paul Meier. "Nonparametric Estimation from Incomplete Observations." In: *Journal of the American Statistical Association* 53.282 (1958), pp. 457–481.

[57] Wolfgang Kerber. "Digital Markets, Data, and Privacy: Competition Law, Consumer Law, and Data Protection." In: *Journal of Intellectual Property Law & Practice* 11.11 (2016), pp. 856–866.

[58] Steffen Klee, Alexandros Roussos, Max Maass, and Matthias Hollick. "NFCGate: Opening the Door for NFC Security Research with a Smartphone-Based Toolkit." In: *14th USENIX Workshop on Offensive Technologies – WOOT '20*. 2020.

[59] John P Klein, Brent Logan, Mette Harhoff, and Per Kragh Andersen. "Analyzing survival curves at a fixed point in time." In: *Statistics in Medicine* 26.24 (2007), pp. 4505–4519.

[60] Marc Kührer, Thomas Hupperich, Christian Rossow, and Thorsten Holz. "Exit from Hell? Reducing the Impact of Amplification DDoS Attacks." In: *23th USENIX Security Symposium – USENIX Security '14*. 2014, pp. 111–125.

[61] Robert H Lande. "The Microsoft-Yahoo Merger: Yes, Privacy is an Antitrust Concern." In: *FTC: Watch* (2008).

[62] Tobias Lauinger, Abdelberi Chaabane, Sajjad Arshad, William Robertson, Christo Wilson, and Engin Kirda. "Thou Shalt Not Depend on Me: Analysing the Use of Outdated JavaScript Libraries on the Web." In: *Proceedings of the 2017 Network and Distributed System Security Symposium – NDSS '17* (2017).

[63] Ben Laurie, Adam Langley, and Emilia Kasper. *Certificate Transparency*. RFC 6962. RFC Editor, 2013. URL: https://www.rfc-editor.org/rfc/rfc6962.txt.

[64] Victor Le Pochat, Tom Van Goethem, Samaneh Tajalizadehkhoob, Maciej Korczynski, and Wouter Joosen. "Tranco: A Research-Oriented Top Sites Ranking Hardened Against Manipulation." In: *Proceedings of the 2019 Network and Distributed System Security Symposium – NDSS '19*. 2019.

[65] Matthew Lee and Jillian Kohler. "Benchmarking and transparency: Incentives for the pharmaceutical Industry's Corporate Social Responsibility." In: *Journal of Business Ethics* 95.4 (2010), pp. 641–658.

[66] *LG Dresden, Urteil v. 11.1.2019 – 1a O 1582/18*. URL: https://dejure.org/2019,16972 (visited on 04/16/2020).

[67] Frank Li. "Remedying Security Concerns at an Internet Scale." PhD thesis. UC Berkeley, 2019.

[68] Frank Li, Zakir Durumeric, Jakub Czyz, Mohammad Karami, Michael Bailey, Damon Mccoy, Stefan Savage, and Vern Paxson. "You've Got Vulnerability: Exploring Effective Vulnerability Notifications." In: *25th USENIX Security Symposium – USENIX Security '16*. 2016.

[69] Frank Li, Grant Ho, Eric Kuan, Yuan Niu, Lucas Ballard, Kurt Thomas, Elie Bursztein, and Vern Paxson. "Remedying Web Hijacking: Notification Effectiveness and Webmaster Comprehension." In: *Proceedings of the 25th International Conference on World Wide Web – WWW '16*. 2016, pp. 1009–1019.

[70] Chaoyi Lu, Baojun Liu, Yiming Zhang, Zhou Li, Fenglu Zhang, Haixin Duan, Ying Liu, Joann Qiongna Chen, Jinjin Liang, Zaifeng Zhang, et al. "From WHOIS to WHOWAS: A Large-Scale Measurement Study of Domain Registration Privacy under the GDPR." In: *Proceedings of the 2021 Network and Distributed System Security Symposium – NDSS '21*. 2021.

[71] Max Maass, Anne Laubach, and Dominik Herrmann. "PrivacyScore: Analyse von Webseiten auf Sicherheits- und Privatheitsprobleme - Konzept und rechtliche Zulässigkeit." In: *INFORMATIK 2017 Workshop "Recht und Technik"*. 2017.

[72] Max Maass, Henning Pridöhl, Dominik Herrmann, and Matthias Hollick. "Best Practices for Notification Studies for Security and Privacy Issues on the Internet." In: *3rd International Workshop on Information Security Methodology and Replication Studies (IWSMR '21)*. 2021.

[73] Max Maass, Alina Stöver, Henning Pridöhl, Sebastian Bretthauer, Dominik Herrmann, Matthias Hollick, and Indra Spiecker. *Supplementary Material for "Effective Notification Campaigns on the Web: A Matter of Trust, Framing, and Support"*. 2020. DOI: 10.5281/zenodo.4075131.

[74] Moxie Marlinspike. "More tricks for defeating SSL in practice." In: *Black Hat USA* (2009).

[75] Matthias Marx, Ephraim Zimer, Tobias Mueller, Maximilian Blochberger, and Hannes Federrath. "Hashing of personally identifiable information is not sufficient." In: *Sicherheit 2018*. 2018, pp. 55–68.

[76] Arunesh Mathur, Gunes Acar, Michael J Friedman, Elena Lucherini, Jonathan Mayer, Marshini Chetty, and Arvind Narayanan. "Dark Patterns at Scale." In: *Proceedings of the ACM on Human-Computer Interaction*. Vol. 3. 2019, pp. 1–32.

[77] Wilfried Mayer, Aaron Zauner, Martin Schmiedecker, and Markus Huber. "No Need for Black Chambers: Testing TLS in the E-mail Ecosystem at Large." In: *Proceedings of the 11th International Conference on Availability, Reliability and Security – ARES '16*. 2016.

[78] Michael L Maynard. "Policing transnational commerce: Global awareness in the margins of morality." In: *Journal of Business Ethics* 30.1 (2001), pp. 17–27.

[79] Philipp Mayring. "Qualitative Content Analysis." In: *Forum Qualitative Sozialforschung* 1.2 (2000).

[80] Aleecia M McDonald and Lorrie Faith Cranor. "The cost of reading privacy policies." In: *I/S: A Journal of Law and Policy for the Information Society* 4 (2008), p. 543.

[81] Manfred Milinski, Dirk Semmann, and Hans-Jürgen Krambeck. "Reputation helps solve the 'tragedy of the commons'." In: *Nature* 415 (2002).

[82] Tobias Mueller, Matthias Marx, Henning Pridöhl, Pascal Wichmann, and Dominik Herrmann. "Sicherheit und Privatheit auf deutschen Hochschulwebseiten: Eine Analyse mit PrivacyScore." In: *25. DFN-Konferenz "Sicherheit in vernetzten Systemen"* (2018).

[83] Maureen K Ohlhausen and Alexander P Okuliar. "Competition, Consumer Protection, and the Right [Approach] To Privacy." In: *Antitrust Law Journal* 80.1 (2015), pp. 121–156.

[84] Ehimare Okoyomon, Nikita Samarin, Primal Wijesekera, Amit Elazari Bar On, Narseo Vallina-Rodriguez, Irwin Reyes, Álvaro Feal, and Serge Egelman. "On The Ridiculousness of Notice and Consent: Contradictions in App Privacy Policies." In: *Proceedings of the Workshop on Technology and Consumer Protection – ConPro '19*. 2019.

[85] Denise Linda Parris, Jennifer L Dapko, Richard Wade Arnold, and Danny Arnold. "Exploring transparency: a new framework for responsible business management." In: *Management Decision* 54.1 (2016), pp. 222–247.

[86] Margaret Sullivan Pepe and Thomas R Fleming. "Weighted Kaplan-Meier Statistics: A Class of Distance Tests for Censored Survival Data." In: *Biometrics* 45.2 (1989), p. 497.

[87] Henning Pridöhl, Pascal Wichmann, Dominik Herrmann, Max Maass, Martin Müller, and Malte. *PrivacyScore/privacyscanner*. 2019. DOI: 10.5281/zenodo.2555037.

[88] Catharine H Rankin et al. "Habituation revisited: an updated and revised description of the behavioral characteristics of habituation." In: *Neurobiology of learning and memory* 92.2 (2009), pp. 135–138.

[89] Justin M Rao and David H Reiley. "The Economics of Spam." In: *Journal of Economic Perspectives* 26.3 (2012), pp. 87–110.

[90] Jingjing Ren, Martina Lindorfer, Daniel J Dubois, Ashwin Rao, David Choffnes, and Narseo Vallina-Rodriguez. "Bug Fixes, Improvements, ... and Privacy Leaks - A Longitudinal Study of PII Leaks Across Android App Versions." In: *Proceedings of the 2018 Network and Distributed System Security Symposium – NDSS '18*. 2018.

[91] Travis Reynolds, Byron T Murray, Jane Kolodinsky, and Jillian Howell. "Contrasting self-reported willingness to pay and demonstrated purchase behavior for energy-saving technologies in a small island developing state." In: *Energy for Sustainable Development* 27 (2015), pp. 18–27.

[92] Franziska Roesner. "Security and Privacy for Untrusted Applications in Modern and Emerging Client Platforms." PhD thesis. 2014.

[93] Lorien Sabatino and Geza Sapi. "Online privacy and market structure: Theory and evidence." In: *DICE Discussion Paper* 308 (2019).

[94] Quirin Scheitle, Oliver Hohlfeld, Julien Gamba, Jonas Jelten, Torsten Zimmermann, Stephen D Strowes, and Narseo Vallina-Rodriguez. "A Long Way to the Top: Significance, Structure and Stability of Internet Top Lists." In: *Proceedings of the 2018 Internet Measurement Conference – IMC '18*. 2018, pp. 478–493.

[95] Yun Shen, Pierre-Antoine Vervier, and Gianluca Stringhini. "Understanding Worldwide Private Information Collection on Android." In: *ArXiv preprint*. 2021. arXiv: 2102.12869.

[96] Wissem Soussi, Marciej Korczyński, Sourena Maroofi, and Andrzej Duda. "Feasibility of Large-Scale Vulnerability Notifications after GDPR." In: *IEEE European Symposium on Security and Privacy 2019 Workshops – EuroS&P '20 Workshops*. 2020.

[97] Ana Pop Stefanija and Jo Pierson. "Practical AI Transparency: Revealing Datafication and Algorithmic Identities." In: *Journal of Digital Social Research* 2.3 (2020), pp. 84–125.

[98] Ben Stock, Giancarlo Pellegrino, Frank Li, Michael Backes, and Christian Rossow. "Didn't You Hear Me? Towards More Successful Web Vulnerability Notifications." In: *Proceedings of the 2018 Network and Distributed System Security Symposium – NDSS '18*. 2018.

[99] Ben Stock, Giancarlo Pellegrino, Christian Rossow, Martin Johns, and Michael Backes. "Hey, You Have a Problem: On the Feasibility of Large-Scale Web Vulnerability Notification." In: *25th USENIX Security Symposium – USENIX Security '16*. 2016, pp. 1015–1032.

[100] Cynthia Stohl, Michael Stohl, and Paul M Leonardi. "Managing Opacity: Information visibility and the paradox of transparency in the digital age." In: *International Journal of Communication Systems* 10.0 (2016), p. 15.

[101] Milan Stute, Sashank Narain, Alex Mariotto, Alexander Heinrich, David Kreitschmann, Guevara Noubir, and Matthias Hollick. "A billion open interfaces for Eve and Mallory: MitM, DoS, and tracking attacks on iOS and macOS through Apple Wireless Direct Link." In: *28th USENIX Security Symposium – USENIX Security '19*. 2019.

[102] Baohong Sun and Vicki G Morwitz. "Stated intentions and purchase behavior: A unified model." In: *International Journal of Research in Marketing* 27.4 (2010), pp. 356–366.

[103] Erik Sy, Christian Burkert, Hannes Federrath, and Mathias Fischer. "Tracking Users across the Web via TLS Session Resumption." In: *Proceedings of the 34th Annual Computer Security Applications Conference – ACSAC '18*. 2018, pp. 289–299.

[104] Qian Tang, Leigh L Linden, John S Quarterman, and Andrew B Whinston. "Improving Internet Security Through Information Disclosure: A Field Quasi-Experiment." In: *Proceedings of the Workshop on the Economics of Information Security – WEIS '13*. 2013.

[105] Qian Tang and Andrew B Whinston. "Do Reputational Sanctions Deter Negligence in Information Security Management? A Field Quasi-Experiment." In: *Production and Operations Management* 29.2 (2020).

[106] Fabian Ullrich, Jiska Classen, Johannes Eger, and Matthias Hollick. "Vacuums in the cloud: analyzing security in a hardened IoT ecosystem." In: *13th USENIX Workshop on Offensive Technologies – WOOT '19*. 2019.

[107] Alexander Ulrich, Ralph Holz, Peter Hauck, and Georg Carle. "Investigating the OpenPGP Web of Trust." In: *Proceedings of the 2011 European Symposium on Research in Computer Security – ESORICS 2011*. 2011, pp. 489–507.

[108] Marie Vasek and Tyler Moore. "Do Malware Reports Expedite Cleanup? An Experimental Study." In: *Workshop on Cyber Security Experimentation and Test*. 2012, pp. 1–8.

[109] Yash Vekaria, Vibhor Agarwal, Pushkal Agarwal, Sangeeta Mahapatra, Sakthi Balan Muthiah, Nishanth Sastry, and Nicolas Kourtellis. "Differential Tracking Across Topical Webpages of Indian News Media." In: *ArXiv preprint*. 2021. eprint: arXiv:2103.04442.

[110] Viswanath Venkatesh, Susan A Brown, and Hillol Bala. "Bridging the Qualitative-Quantitative Divide: Guidelines for Conducting Mixed Methods Research in Information Systems." In: *The Mississippi Quarterly* 37.1 (2013), pp. 21–54.

[111] Jessica Vitak, Katie Shilton, and Zahra Ashktorab. "Beyond the Belmont Principles: Ethical Challenges, Practices, and Beliefs in the Online Data Research Community." In: *Proceedings of the 19th ACM Conference on Computer-Supported Cooperative Work & Social Computing*. 2016.

[112] David Weil, Archon Fung, Mary Graham, and Elena Fagotto. "The effectiveness of regulatory disclosure policies." In: *Journal of policy analysis and management* 25.1 (2006), pp. 155–181.

[113] Haitao Xu, Shuai Hao, Alparslan Sari, and Haining Wang. "Privacy Risk Assessment on Email Tracking." In: *Proceedings of the 2018 IEEE Conference on Computer Communications – INFOCOM '18*. 2018, pp. 2519–2527.

[114] Zhonghao Yu, Sam Macbeth, Konark Modi, and Josep M Pujol. "Tracking the Trackers." In: *Proceedings of the 25th International Conference on World Wide Web – WWW '16*. 2016, pp. 121–132.

[115] Nicholas Zarimpas, ed. *Transparency in nuclear warheads and materials*. SIPRI Monographs. London, England: Oxford University Press, 2003. ISBN: 9780199252428.

[116] Jay Zarnikau. "Consumer demand for 'green power' and energy efficiency." In: *Energy policy* 31.15 (2003), pp. 1661–1672.

[117] Eric Zeng, Frank Li, Emily Stark, and Adrienne Porter Felt. "Fixing HTTPS Misconfigurations at Scale: An Experiment with Security Notifications." In: *Proceedings of the 18th Annual Workshop on the Economics of Information Security – WEIS '19*. 2019.

[118] Yunhui Zhuang, Yunsik Choi, Shu He, Alvin Chung-Man Leung, Gene Moo Lee, and Andrew B Whinston. "Information Disclosure and Security Vulnerability Awareness: A Large-Scale Randomized Field Experiment in Pan-Asia." In: *Hawaii International Conference on System Sciences*. 2020.

ERKLÄRUNG ZUR DISSERTATIONSSCHRIFT

gemäß § 9 der Allgemeinen Bestimmungen der Promotionsordnung der Technische Universität Darmstadt vom 12. Januar 1990 (ABI. 1990, S. 658) in der Fassung der 8. Novelle vom 1. März 2018

Hiermit versichere ich, Max Jakob Maaß, die vorliegende Dissertationsschrift ohne Hilfe Dritter und nur mit den angegebenen Quellen und Hilfsmitteln angefertigt zu haben. Alle Stellen, die Quellen entnommen wurden, sind als solche kenntlich gemacht worden. Eigenzitate aus vorausgehenden wissenschaftlichen Veröffentlichungen werden in Anlehnung an die Hinweise des Promotionsausschusses Fachbereich Informatik zum Thema „Eigenzitate in wissenschaftlichen Arbeiten" (EZ-2014/10) in Kapitel „*Collaborations and My Contribution*" auf Seiten xxi bis xxii gelistet. Diese Arbeit hat in gleicher oder ähnlicher Form noch keiner Prüfungsbehörde vorgelegen. In der abgegebenen Dissertationsschrift stimmen die schriftliche und die elektronische Fassung überein.

Darmstadt, 21. Mai 2021

Max Jakob Maaß