KU LEUVEN

ARENBERG DOCTORAL SCHOOL
Faculty of Engineering Science

Sound data sets and methods for web security research

Victor Le Pochat

Supervisors:
Prof. dr. ir. Wouter Joosen
Dr. ir. Lieven Desmet

Dissertation presented in partial fulfillment of the requirements for the degree of Doctor of Engineering Science (PhD): Computer Science

May 2023

Sound data sets and methods for web security research

Victor LE POCHAT

Examination committee:
Prof. dr. ir. Dirk Vandermeulen, chair
Prof. dr. ir. Wouter Joosen, supervisor
Dr. ir. Lieven Desmet, supervisor
Prof. dr. ir. Frank Piessens
Prof. dr. Katrien Verbert
Prof. dr. Michel van Eeten
(Technische Universiteit Delft)
Prof. dr. Maciej Korczyński
(Université Grenoble Alpes)

Dissertation presented in partial fulfillment of the requirements for the degree of Doctor of Engineering Science (PhD): Computer Science

May 2023

Uitgegeven in eigen beheer, Victor Le Pochat, Celestijnenlaan 200A bus 2402, B-3001 Leuven (Belgium)

Acknowledgments

I thank my advisor Wouter for the opportunity to join DistriNet and for giving me the freedom and independence to do research. I would like to thank Lieven, Frank, Katrien, Maciej, and Michel for agreeing to be part of my examination committee, and prof. Vandermeulen for chairing the defense, therefore playing an important role in achieving this milestone. Thank you Tom, for teaching me the ropes of measuring the web and writing it all up. Thank you to the WebSecPriv and SIG Analytics teams for the valuable discussions and exchange of ideas.

A great highlight of my PhD has been the many collaborations, enabling me to work together with and learn from fellow researchers on a great variety of projects. I would like to thank all my collaborators for enabling such a great experience. I want to thank Maciej for hosting me in Grenoble, and Toby and Damon for hosting me in New York.

Thank you to Katrien, Annick, Annelies, and An at the DistriNet business office, all the past and current members of the departmental secretariat and systems group, and Lieve as the departmental manager, for ensuring operations run smoothly. Also my thanks to Emmanuel at the press office for helping to spread the word about our research.

The research conducted during the doctoral programme was supported by a PhD fellowship from the Research Foundation Flanders - FWO (11A3419N, 11A3421N) and travel grants (V426920N, K143322N); the Erasmus+ programme; the Research Fund KU Leuven; and the Flemish Research Programme Cybersecurity.

Abstract

Sound research practices form the foundation of valid, reliable, and trustworthy research results. In the context of web security research, the importance of measurements that yield the empirical real-world data used for analyzing and improving security on the web emphasizes the need to use sound data sets and methods that allow these measurements to be accurate, comprehensive, representative, and transparent. In this dissertation, we discuss three case studies in web security that have affinity with the discipline of meta-research, which critically evaluates research practices and proposes new methods to improve and refine the way in which research is conducted. The three presented case studies contribute to critically analyzing current web security data sets or systems that are commonly held to be reliable, while we also propose improved methods as well as discuss data set considerations.

In the first part of the dissertation, we present our first case study, analyzing and improving rankings of the most popular websites or domain names on the Internet. These rankings form an important data source for many web security, privacy, and Internet measurement studies. We show how previously commonly used rankings hold potentially undesirable properties that endanger the soundness, validity, and reproducibility of research. We also propose the novel Tranco ranking that improves upon these properties. Tranco combines existing rankings transparently, aggregates across a 30-day period by default to improve long-term stability, and the resulting ranking is made available in a reproducible manner. We confirm with a long-term evaluation that Tranco better matches the properties desirable for research usage.

In the second part of the dissertation, we present two case studies on large-scale automated decision-making systems. These are seen as essential tools for processing security-related decisions at scale, and are commonly deployed to handle critical security tasks. However, there are concerns that they are ineffective at this task, potentially endangering security on the web. In our second case study, we develop a hybrid approach to resolve collisions between benign and malicious domains generated and used by the Avalanche botnet. As erroneous law enforcement decisions would result in unjustified website takedowns and the risk of the botnet reemerging respectively, we involve a human investigator for those domains where an automated model is least certain. This approach reduces the errors that result from blind trust in the automated decision-making system. In our third

case study, we audit Facebook's enforcement of its self-developed policy on political ads. We find that even simple rules for detecting violating ads are not implemented, while many benign ads are falsely taken down, suggesting Facebook's enforcement is imprecise. Our audit reveals the limitations of large-scale automated decision-making systems and questions their appropriateness for security problems with important societal impact.

We conclude the dissertation with closing remarks on enablers and challenges for web security research, focusing on the importance of data sets for analyzing security issues and ecosystems, and for developing improved security solutions. We also provide an outlook on future research topics that explore remaining gaps in the current state of the art, in domain rankings and large-scale web measurements in general, as well as automated decision-making systems. Further work in this domain helps to enable more complete and thorough insights into malicious online practices, allowing us to develop better solutions that make the web a more secure place for all.

Beknopte samenvatting

Goede onderzoekspraktijken vormen de basis voor geldige, betrouwbare en geloofwaardige onderzoeksresultaten. In de context van webbeveiligingsonderzoek benadrukt het belang van metingen die de empirische gegevens leveren voor het analyseren en verbeteren van beveiliging op het web de noodzaak om deugdelijke datasets en methoden te gebruiken die het mogelijk maken om deze metingen nauwkeurig, volledig, representatief en transparant uit te voeren. In dit proefschrift bespreken we drie gevalstudies in webbeveiliging die verwant zijn aan de discipline van meta-onderzoek, waarin deze onderzoekspraktijken kritisch worden geëvalueerd en nieuwe methoden worden voorgesteld om de manier waarop onderzoek wordt uitgevoerd te verbeteren en te verfijnen. De drie voorgestelde gevalstudies dragen bij aan een kritische analyse van huidige datasets of systemen voor webbeveiliging die algemeen als betrouwbaar worden beschouwd, terwijl we ook verbeterde methoden voorstellen naast het bespreken van overwegingen met betrekking tot datasets.

In het eerste deel van het proefschrift presenteren wij onze eerste studie, rond het analyseren en verbeteren van ranglijsten van de populairste websites of domeinnamen op het internet. Deze ranglijsten vormen een belangrijke gegevensbron voor veel studies in webbeveiliging, privacy en internetmetingen. Wij tonen hoe de eerder vaak gebruikte ranglijsten potentieel ongewenste eigenschappen bezitten die de deugdelijkheid, geldigheid en reproduceerbaarheid van onderzoek in gevaar brengen. Wij stellen ook de nieuwe Tranco-ranglijst voor die deze eigenschappen verbetert. Tranco combineert bestaande ranglijsten op transparante wijze en aggregeert standaard over een periode van 30 dagen om de stabiliteit op lange termijn te verbeteren. De resulterende ranglijst wordt op reproduceerbare wijze beschikbaar gesteld. Wij bevestigen met een langetermijnevaluatie dat Tranco beter voldoet aan de eigenschappen die wenselijk zijn voor onderzoeksgebruik.

In het tweede deel van het proefschrift presenteren wij twee studies over grootschalige geautomatiseerde besluitvormingssystemen. Deze worden gezien als essentiële instrumenten voor het verwerken van beveiligingsverwante beslissingen op schaal, en worden vaak ingezet om kritieke beveiligingstaken af te handelen. Er bestaat echter bezorgdheid dat zij deze taak niet doeltreffend uitvoeren, waardoor de veiligheid op het web in gevaar kan komen. In onze tweede studie ontwikkelen we een hybride aanpak om

conflicten op te lossen tussen goedaardige en kwaadaardige domeinen die door het Avalanche-botnet worden gegenereerd en gebruikt. Aangezien foutieve beslissingen zouden leiden tot respectievelijk het ongerechtvaardigd neerhalen van websites en het risico dat het botnet opnieuw opduikt, schakelen wij een menselijke onderzoeker in voor die domeinnamen waar het geautomatiseerde model het minst zeker is. Deze aanpak vermindert de fouten die het gevolg zijn van blind vertrouwen in het geautomatiseerde besluitvormingssysteem. In onze derde studie auditeren wij de handhaving door Facebook van hun zelfontwikkelde beleid inzake politieke advertenties. Wij vinden dat zelfs eenvoudige regels voor het opsporen van advertenties die in strijd zijn met het beleid niet worden toegepast, terwijl veel goedaardige advertenties ten onrechte worden verwijderd, wat erop wijst dat Facebooks handhaving onnauwkeurig is. Ons onderzoek onthult de beperkingen van grootschalige geautomatiseerde besluitvormingssystemen en stelt de geschiktheid ervan voor veiligheidsproblemen met een belangrijke maatschappelijke impact in vraag.

We sluiten het proefschrift af met slotopmerkingen over de elementen die webbeveiligingsonderzoek mogelijk maken net als de uitdagingen die overblijven, met de nadruk op het belang van datasets voor het analyseren van beveiligingsproblemen en ecosystemen, en het ontwikkelen van verbeterde beveiligingsoplossingen. We blikken ook vooruit op toekomstige onderzoeksonderwerpen die hiaten in de huidige stand van de techniek verkennen, in domeinranglijsten en grootschalige webmetingen in het algemeen, en geautomatiseerde besluitvormingssystemen. Verder werk in dit domein helpt bij te dragen aan een vollediger en grondiger inzicht in kwaadaardige online praktijken, waardoor we betere oplossingen kunnen ontwikkelen die het web voor iedereen veiliger maken.

Contents

Abstract iii

Beknopte samenvatting v

Contents vii

1 Introduction 1
- 1.1 Sound web security research . . . 1
  - 1.1.1 Meta-research in web security . . . 3
- 1.2 Dissertation outline . . . 14

I Analyzing and improving domain rankings 17

2 Prologue 19
- 2.1 Domain rankings . . . 19
- 2.2 Usage and impact of Tranco . . . 21
- 2.3 Overview . . . 24

3 Tranco: A Research-Oriented Top Sites Ranking Hardened Against Manipulation 27
- 3.1 Introduction . . . 28
- 3.2 Methodology of top websites rankings . . . 29
  - 3.2.1 Alexa . . . 30
  - 3.2.2 Cisco Umbrella . . . 31
  - 3.2.3 Majestic . . . 31
  - 3.2.4 Quantcast . . . 31
- 3.3 Quantitative comparison . . . 32
  - 3.3.1 Similarity . . . 32
  - 3.3.2 Stability . . . 33
  - 3.3.3 Representativeness . . . 33
  - 3.3.4 Responsiveness . . . 34
  - 3.3.5 Benignness . . . 35

3.4 Usage in security research . . . 36
3.4.1 Survey and classification of list usage . . . 36
3.4.2 Influence on security studies . . . 37
3.5 Feasibility of large-scale manipulation . . . 39
3.5.1 Alexa . . . 40
3.5.2 Cisco Umbrella . . . 44
3.5.3 Majestic . . . 46
3.5.4 Quantcast . . . 49
3.6 An improved top websites ranking . . . 51
3.6.1 Defending existing rankings against manipulation . . . 52
3.6.2 Creating rankings suitable for research . . . 53
3.7 Related work . . . 56
3.8 Conclusion . . . 57

4 Evaluating the Long-term Effects of Parameters on the Characteristics of the Tranco Top Sites Ranking 59
4.1 Introduction . . . 60
4.2 Methods of the Tranco ranking . . . 61
4.3 Analysis of Tranco's properties . . . 62
4.3.1 Similarity with component rankings . . . 62
4.3.2 Comparison with web traffic . . . 63
4.3.3 Stability over time . . . 64
4.3.4 Responsiveness . . . 66
4.3.5 Benignness . . . 70
4.3.6 Anomalies . . . 71
4.3.7 Combination method . . . 72
4.3.8 Structure . . . 73
4.4 Related work . . . 73
4.5 Conclusion . . . 74

II Auditing automated decision-making systems 77

5 Prologue 79

6 A Practical Approach for Taking Down Avalanche Botnets Under Real-World Constraints 87
6.1 Introduction . . . 88
6.2 Background . . . 90
6.2.1 Domain generation algorithms . . . 90
6.2.2 Taking down the Avalanche infrastructure . . . 91
6.3 Problem statement . . . 92
6.3.1 Making accurate takedown decisions . . . 92
6.3.2 Constraints for distinguishing malicious and benign domains . . . 93
6.3.3 Ground truth data . . . 96

6.3.4 Ethical considerations . . . 97
6.4 Data set analysis and feature extraction . . . 97
6.4.1 Life cycle of a domain . . . 97
6.4.2 General insights . . . 99
6.4.3 Summary of feature sets . . . 101
6.4.4 Omitted features . . . 103
6.5 Analysis of machine learning-based classification . . . 105
6.5.1 Experimental protocol . . . 105
6.5.2 Results . . . 107
6.6 Discussion . . . 112
6.6.1 Evasion . . . 113
6.6.2 Availability of data sets . . . 114
6.7 Related work . . . 115
6.8 Conclusion . . . 117
6.A Machine learning protocol . . . 117
6.B Evaluation of machine learning algorithms . . . 119

7 An Audit of Facebook's Political Ad Policy Enforcement 121
7.1 Introduction . . . 122
7.2 Background . . . 124
7.2.1 Political ad policy . . . 124
7.2.2 Policy enforcement . . . 125
7.2.3 Transparency tools . . . 126
7.2.4 Related work . . . 126
7.3 Enforcement Errors and Their Impact . . . 127
7.4 Data Collection . . . 128
7.4.1 Scope and method . . . 129
7.4.2 Data set description . . . 129
7.4.3 Data set validation . . . 131
7.4.4 Ethics . . . 133
7.4.5 Limitations . . . 133
7.5 Ad-level Enforcement . . . 134
7.5.1 Current ad policy enforcement . . . 135
7.5.2 Ads incorrectly detected as political . . . 136
7.6 Page-level Enforcement . . . 138
7.6.1 Reaction to enforcement . . . 138
7.6.2 Current enforcement by advertiser class . . . 141
7.6.3 Missed ads by political advertisers . . . 144
7.7 Discussion . . . 147
7.8 Conclusion . . . 150
7.A Ad Library web portal . . . 150
7.B Data collection timeline . . . 152
7.C Discarded pages . . . 152
7.D Legal framework for online political and issue advertising . . . 154
7.E Topic codebook . . . 154

7.F Page categories 155
7.G Ad Library report dates 158

III Conclusion **161**

8 Conclusion **163**
8.1 The importance of data sets 163
8.2 Outlook and future work 165
8.2.1 Domain rankings 165
8.2.2 Large-scale web measurements 167
8.2.3 Automated decision-making systems 168
8.3 Closing thoughts 170

Bibliography **171**

1 Introduction

1.1 Sound web security research

The web is by now indispensable in our daily lives. It is gaining ever-increasing importance, controlling many or often nearly all processes in domains such as business, finance, government, infrastructure, entertainment, or social networking. The scale of the web continues expanding: in raw numbers, there are already over 350 million registered domain names [494], and new services and tools become available day after day. Its coverage also grows, with nearly the entire world population having access to the Internet, and the web being accessed by more cultures, audiences, networks, or devices.

Coupled with this continuing growth, the security issues that plague the web become more prevalent and impactful. With the web and its attack surface growing, cyber attacks are steadily on the rise, and with the web playing an ever more important role, these attacks can have a severe impact on essential infrastructure such as power grids [273]. The domain of web security research then examines these online interactions where adversaries target web browsing and web applications [29], with attacks ranging from exploiting vulnerabilities in the protocols and tools in web browsing to steal user credentials or personal data, to engaging in online abuse and cybercrime to defraud users. The goal of our research is to study these security issues, understand the modi operandi of the attackers and defenders, and design solutions that mitigate attacks and help to better protect web users against current and emerging threats.

Central to most web security research studies are *measurements* across a (large) set of websites and other Internet properties, which yield the empirical real-world data that allow us to understand security on the web, which is analyzed and translated into research results, insights, and the development of potential countermeasures. With the web itself becoming more complex and attackers deploying more sophisticated operations that lure unsuspecting users while seeking to remain hidden from defenders, the techniques necessary to study web security also need to be more elaborate and well-designed in order

to keep up with this increase in complexity. A large variety of measurement techniques is also being developed to account for this diversity in web security research areas [390].

When we develop and use measurements to evaluate security on the web, we want to be certain that our research is as valid and sound as possible, capturing the real-world state of the web as it is. This is not only important to ensure that our analyses and findings reflect the actual security issues that occur across the web, but also to ensure that the mitigations that we propose are actually effective at having a positive effect on web security and preventing people from being harmed by these security issues. To ensure that we can execute web security measurements and the research in general in a valid and sound manner, there are several underlying principles that can guide us.

- We want our measurements to be *accurate*. This means using data sets and methods that are free of any errors or biases, e.g., selecting inappropriate measurement parameters that could change how the measurement is conducted, as these errors could propagate to the research results [291, 486], possibly leading us to misinterpret the importance of security issues on the web today. We also want to ensure that there are no false assumptions about these data sets and methods. For example, rankings of popular websites are sometimes considered to contain only benign websites and are then used as such in classifiers of benign and malicious websites. However, these rankings actually also contain malicious websites from time to time – if an attacker succeeds in having many users or machines access a malicious website, it is by definition popular –, which means that any classifiers that are trained upon those rankings could actually be less effective at detecting malicious websites given the wrong data that they were provided with.
- We want our measurements to be *comprehensive*. This encompasses that the measurements are large-scale: by now, crawling a million websites – e.g., to measure the prevalence of a web security vulnerability on a large sample of the web – is considered a relatively standard practice [163], while researchers routinely measure the entire IPv4 address space [151] and advancements are being made to explore more of the IPv6 address space [194]. Measurements should also cover a variety of vantage points, countries, networks, demographics, languages, and so on, to ensure that they maximally observe security issues for all populations and services. For example, studies into domain squatting may need to account for international audiences, as these may be differently affected by this web security issue [288, 290]. This should also enable any mitigations that are developed to fairly protect and be beneficial for all web users. For example, websites enforce HTTPS inconsistently between geographic locations, which leads to unequal user protection across the world, and then becomes exploitable by an attacker through redirecting traffic between regions [30].
- We want our measurements to be *representative*. They should reflect how the web is typically used and what security issues this might affect, taking into account that the web can be used in different ways: while browsing patterns from human users tend to concentrate on a small number of popular websites (with meaningful

content) [414], machines and background API traffic may more heavily rely on infrastructural domains such as CDNs, where other (and possibly more invisible) web security issues may be more important. Similarly, web security postures may differ between desktop and mobile websites and therefore users [487].

- We want our measurements to be *transparent*. This includes using or providing open methods and data sets, as well as clearly describing how these were used. This allows others to evaluate whether the research was done using appropriate methods and data sets, but also enables them to reproduce the research or to build upon this prior work and accelerate their own research.

1.1.1 Meta-research in web security

A critical reflection on the aforementioned principles for web security measurements can be framed within the discipline of *meta-research* (or *metascience*). This discipline addresses the critical evaluation of the various methods and practices in scientific research in general [249], i.e., 'does research on research'. Its goal is to understand whether current research is sufficiently sound and reliable, and develop new best practices to improve and refine the way in which research is conducted. This supports enabling research findings and claims to be credible and trustworthy, allowing to build upon them for making informed decisions, or for example within the scope of security to use them for developing better countermeasures against attacks. One of the main underlying topics is understanding the range of biases that may emerge when doing research, as they negatively impact the aforementioned principles such as correctness or soundness, and searching mitigation strategies for these biases [106, 209, 249].

Ioannidis et al. [249] introduced a categorization of meta-research for the different phases of the research cycle. This categorization captures five areas of meta-research:

- the *methods* used when designing and conducting studies. This includes the development of sound data collection or the use of appropriate data sets. The analysis of methods seeks to account for a variety of biases from flawed methods and instruments, selection (e.g., sampling) biases, to inappropriate statistical analyses. In web security, this could for example be due to web crawling strategies that fail to account for detection of automated crawling, or using non-representative lists of websites. These might give a wrong impression of the reliability or accuracy of findings or solutions based upon them.
- *reporting* or communicating research, also avoiding biases due to misinterpretation of results, with a risk of the wrong conclusions being drawn if the research results are not accurately and completely conveyed.
- *reproducibility* of research, allowing others to verify research, and avoiding biases from one-off observations and instead allowing to understand if the research captures a genuine trend. In web security research, this involves, e.g., accurately

describing which data sets and tools were used to conduct a measurement, and which assumptions were made when developing an evaluation or countermeasure.

- *evaluation* of research, primarily corresponding to the peer review process, which might suffer from biases such as favoring positive results or randomness in paper acceptance.
- *incentives* for research, or an understanding of what research is favored, including perceptions related to metrics for papers (e.g., citations) or funding criteria.

Of these categories, methods and reproducibility tend to be more specific to a given research field, such as web security, but all categories play some role in ensuring soundness and validity within the scientific process. We use this categorization by Ioannidis et al.to gain a better understanding on the (web) security research community's efforts to examine its own research practices. We use the framework to characterize which areas of meta-research are most commonly studied, and which areas receive the most attention in terms of developing and enforcing improved research practices. We also compare these meta-research findings with experiences from another academic community, in this case the Internet measurement community, to understand how these practices may differ between research communities and observe areas where the communities can learn from each other. With our overview, we seek to understand current practices, priorities, or open questions related to the way in which cyber security research is conducted – in a way, we will engage in 'meta-research on meta-research'. Our work is meant as an encouragement for the research community to continue its self-reflective practices, and we hope that it can contribute to these ongoing efforts to improve cyber security research.

To compile our overview of meta-research work related to cyber security, we select topics that are relevant to each category in Ioannidis et al.'s framework and then search the relevant literature on each topic. To discover papers, we do a broad search across Google Scholar and the ACM Digital Library for papers that match topic-related keywords (e.g., "peer review" for the *evaluation* category), iteratively processing the references of discovered papers to compile the final set of papers that we discuss in our overview. We do not set an explicit time range for our survey; we observe that some works already date back 20 years or more. For areas where the relevant literature is broad and extensive, we exemplify the work in the area with papers related to the field of web security and privacy, but seek to extend it to the broader computer security field in general. We also leverage our familiarity with web-related research by selecting the Internet measurement community to compare research practices, as this community addresses similar topics and issues. We conduct our survey of relevant meta-research work in this community in the same way as for the cyber security community.

The remainder of this section is structured as one paragraph per category in Ioannidis et al.'s framework – *methods*, *reporting*, *reproducibility*, *evaluation*, and *incentives* –, ending with a paragraph concluding our overview and examining the trends observed and the lessons we can learn going forward.

Methods Crucial to the validity of research is *conducting* it using the best scientific methods and practices possible. Otherwise, there is a risk that the experiments and their results are not truly representative or accurate.

Data collection is a crucial phase of a research project, as all subsequent analyses and results depend on the accuracy and validity of the acquired data. A common denominator to many studies related to web security, web privacy, and Internet measurement in general is the use of *large-scale measurements* for this data collection. Pour et al. [390] survey the use of various Internet measurement techniques in recent cyber security work, creating a taxonomy based on the type of security issue that was studied. Unsurprisingly, recent work has critically analyzed methods that are regularly used in Internet measurement research, often formulating recommendations for how researchers should use them or proposing improved solutions. Already in 2004, Paxson [379] outlined strategies for sound Internet measurement, such as calibrating measurements, inspecting raw data, and designing for reproducibility. Collecting web data often involves 'crawlers' that scrape and store a web page's contents. Ahmad et al. [27] compared web crawlers with varying technologies and feature sets, finding that the choice of crawler may significantly impact measurements. Zeber et al. [529] compared crawlers with each other and with human-generated traffic, finding that crawling results can vary significantly over time as well as across platforms. Krumnow et al. [279] analyzed how the popular OpenWPM crawling framework is detectable and how its measurements can therefore be prevented or poisoned, which introduces errors into the obtained results. Szurdi et al. [467] found that cybercrime must be measured using multiple vantage points and profiles, with special attention to cloaking, in order to obtain reliable results. Jueckstock et al. [256] measured how the browser configuration and network vantage point cause significant biases for web privacy and security measurements. Demir et al. [141] measured how different experimental setups such as the browser, location, user interaction, and time may significantly influence web measurements. Roth et al. [413] measured how websites have inconsistent security policies between browsing profiles. This also has implications for measurements, as these may misreport findings if website behavior changes between profiles or page accesses. Wan et al. [506] found that Internet scan results depend on their origin, i.e., the location, network type, or protocol. Cassel et al. [107] found that frameworks for emulating mobile browsers on desktop may produce results that differ from real mobile browsing, causing incorrect findings about the mobile web specifically.

Across the board, many domains of cyber security research have seen studies on *best practices and pitfalls*. Given the breadth of our field, we give a non-exhaustive selection of example studies that address these issues. Rossow et al. [412] studied issues in malware research, ranging from incorrect datasets, a lack of transparency on methods or results, unrealistic settings, to a lack of safety procedures for containing the malware. Botacin et al. [96] identified twenty pitfalls in malware research through a literature review, adding issues such as closed data sets. Arp et al. [59] identified ten common pitfalls in the application of machine learning in security research, at all stages of the machine learning workflow. Eberz et al. [152] found that evaluations of behavioral biometric authentication systems failed to report error distributions, which may have led to incorrect evaluations. Sugrim et al. [465] proposed robust metrics for the evaluation of authentication systems

that use machine learning, as they found that existing commonly used metrics were incomplete or hard to compare. Das et al. [137] analyzed how studies use hardware performance counters and whether they acknowledged and/or addressed limitations in using them for security applications. Van der Kouwe et al. [271] analyzed how pitfalls may affect the validity of performance benchmarking in systems security papers, if they cause flaws such as an incomplete evaluation, irrelevant or unsound results, or a lack of reproducibility. Polakis et al. [389] described the various methods used in all phases of a measurement study on social networks, from ethical considerations to data collection and processing techniques.

Data sets form another subject of scrutiny, as there are often questions about reliability and validity, especially if these data sets are difficult to acquire or generated opaquely (e.g., by a commercial third party) [381]. For example, VirusTotal is a commonly used, but commercial source for labeling entities such as files and URLs as benign or malicious. Peng et al. [382] studied how reliable VirusTotal is for detecting phishing websites, finding varying and inadequate detection performance as well as inconsistent labeling. Zhu et al. [538] studied how researchers use VirusTotal to label malware, and analyzed how reliable the data set is in terms of accuracy, independence, and stability over time. More broadly, Feal et al. [175] found that blocklists are opaquely constructed, may be slow to update, may either label records differently or share labels and therefore have high overlap, and are not always well documented. Our works on domain rankings and on domain categorization [486] fall in line with these data set studies, similarly finding opaque methods and disagreement. Researcher-generated data sets may also suffer from a lack of coverage. For example, Cuevas et al. [132] found that scraping-based measurements 'by proxy' on online anonymous marketplaces systematically underestimate metrics such as revenue or the number of discovered listings.

Another example is the tension between using *real-world versus simulated data sets*. Real-world data has the perception of being more accurate and representative, but comes with substantial challenges for data collection and publication, not in the least due to the need to obtain permission to collect data and publish a (usable) anonymized version if the data pertains to human behavior [12]. Simulated data overcomes these issues and better allows for repeatable and comparable security experiments, but the community often questions its validity, as it is difficult to assess the quality and representativeness of generated data [12]. Indeed, problems with simulated data sets are known to exist and significantly affect research results. For example, the data collection strategy affects the perceived performance of website fingerprinting attacks [406], and standard data sets for evaluating intrusion detection systems contain significant noise or even errors that impact attack performance [162, 304].

A particular body of research focuses on methods for studying *usability for security and privacy*, which usually entails collecting data from humans through specific methods (e.g., interviews) and analyzing that data qualitatively as opposed to quantitatively [277]. Fujs et al. [190] surveyed the use of such qualitative methods in security research, finding that interviews are most common. Since the rest of the security community may be unfamiliar with these methods, as their research tends to be quantitative, special care is

taken to show the validity of research results that originate from these qualitative methods. Schechter [424] summarized pitfalls and good practices for describing security and privacy experiments that involve human subjects, including the experiment design and setup but also the reporting on statistical tests. Redmiles et al. [399] compiled guidelines for conducting surveys in security and privacy studies, including how to design the questions, achieve a representative sample of participants, and test the questions upfront. Ortloff et al. [370] examined the process of coding (or labeling) data qualitatively for usable security and privacy studies, recommending that the number of coders should be adapted to the data type. Unfortunately, these best practices do not appear to always be followed. Groß [209] analyzed the reliability of statistical analyses in security user studies, finding systemic issues such as low statistical power that put the validity of the results into questions. They use their findings to provide recommendations for supporting and requiring more reliable studies. Kaur et al. [258] surveyed human factors security research over ten years, finding, a.o., biases in population sampling, and a lack of theorization that should be the result from inductive methods such as grounded theory.

Ethical considerations for conducting research are meant to ensure that no harm is done while studying a security or privacy system. Existing frameworks for ethical review may not be adapted to the needs of the (web) security field. Van der Ham and van Rijswijk-Deij [218] describe the shortcomings of processes involving ethical review boards such as an Institutional Review Board for Internet measurements as these often fall out of those boards' scope, and design an alternative framework with guidelines for ethical measurements. Macnish and van der Ham [313] continue this line for security research ethics, using two case studies of controversial studies to motivate how current methods and guidance are inadequate, as review boards provide insufficient guidance and ethical oversight for practitioners is lacking. It is then often up to the community itself to set their own ethical standards and provide guidelines to researchers. The Menlo Report [69], which outlines the principles of respect for persons, beneficence, justice, and respect for law and public interest, is commonly seen as the main framework for ethical computer science research. Reidsma et al. [400] propose a practical framework for addressing the specificities of cybersecurity research when passing through ethical review boards or designing relevant university policies. Allman and Paxson [44] provide guidelines for ethically sharing data from network measurements, preventing risks such as privacy leaks and setting acceptable use policies including appropriate acknowledgments. Conducting research ethically is increasingly enforced at top-tier security conferences, with measures ranging from mandatory descriptions of the ethical considerations made, to research ethics committees reviewing potentially contentious cases [103]. Zhang et al. [533] surveyed ethical considerations in computer security research, including what ethical requirements conferences impose, how papers discuss ethics, and whether researchers apply ethical practices. They also give recommendations on how to learn about ethical requirements, apply them in practice, and describe them appropriately. Feitelson [176] uses the 2021 controversy on the "Hypocrite Commits" paper, which analyzed developer reaction to intentionally introduced bugs, as a starting point for surveying developers and researchers on what they consider ethically acceptable research practices, formulating recommendations based on the insight that developers are willing

to contribute to research if it is conducted transparently and in good faith. Pauley and McDaniel [377] describe the ethical considerations seen recently in practice in Internet measurement research, finding that this community still lacks a cohesive approach.

Reporting *Communicating* research well is essential for ensuring that it reaches the intended audience(s) without being misinterpreted or misrepresented. A research study and its results can be of interest to multiple stakeholders. Fellow researchers can build upon prior work, relate the findings of prior work to their work, or learn about methods and data sets used. Policymakers can use research results as a foundation for new regulations that seek to improve security and privacy, e.g., by prohibiting privacy-invasive practices that research has found in the wild. Industry companies can integrate state-of-the-art research solutions into their tools or processes to improve their security posture. Finally, researchers can communicate the real-world impact of their findings to the public at large, e.g., directly or through the media, and give actionable guidance such that the public can improve their own security and privacy practices. However, it appears there is little research into how these different forms of *science communication* are used in security research. As one example, Narayanan and Lee [355] reflected on the success of their engagement with policymakers, carriers, journalists, and users for their security policy audit of SIM swapping attacks. Pennekamp et al. [383] proposed a framework for conducting cybersecurity research for industrial applications, and collaboration with companies to enable such interdisciplinary research.

Next to studying research that *is* communicated, there is a concern for research that *is not* being made public, either because its results are negative, deemed insignificant, or deemed undesirable, or because it is kept proprietary. *Publication bias* broadly refers to any bias that may cause specific research to be overrepresented or underrepresented in what is actually published, based on the outcomes of that research [144, 209]. The most commonly regarded form is the omission of negative results, where a hypothesis could not be confirmed nor falsified, or an expected phenomenon was not observed, because researchers are less inclined to submit them for publication, and reviewers and other research gatekeepers (e.g., editors, funders) are less inclined to appreciate them. This causes positive results to be overrepresented, extending to an incentive to always find (statistically significant) results. This may trigger questionable practices such as performing many analyses on data until significant results are found ("*p* hacking"). Not publishing negative results may also mean that other researchers waste time and resources retrying those experiments, only to find (and discard) the negative results. This bias also forms a threat for meta-analyses through literature surveys, as these may erroneously conclude only positive findings, as the negative results that run counter to those findings have simply not been published.

Groß [209] showed empirically that the cyber security user study field suffers from a publication bias, with smaller studies without significant results going unpublished. Such user studies might be among the type of study that is most vulnerable to publication bias, as they heavily rely on statistical inferences across relatively small populations, where there is a higher risk of selectively executing analyses and reporting results that

support a hypothesis as well as reporting results with small effect sizes and low statistical power. In security, publication bias may also be due to potential underreporting of vulnerabilities, where papers are not submitted or published in the first place, for example if the vulnerable entity requests that the publication is delayed or stopped altogether, leading to unreliable aggregate vulnerability statistics [115]. Afterwards, there is also a belief that within the research community, papers presenting attacks are more readily accepted than papers proposing defenses [453], potentially giving an appearance that attacks are more prevalent (if they are allowed to be published, as mentioned above). Boucher and Anderson [97] discuss one example of the difficulties that may emerge in academically publishing a discovered vulnerability, as their public disclosure was used as grounds for paper rejection.

One proposed solution to alleviate some publication bias is *preregistration*, where the intended aim, research questions, hypotheses, methods, data sets, analyses, etc. are established in a document before the actual experiments take place [364]. However, it seems that this practice is very uncommon in security and privacy research, possibly also due to the exploratory or vulnerability-oriented nature of many studies, which does not always allow for a detailed experimental design upfront.

Reproducibility *Verifying* research can be achieved by seeking to reproduce it. Successfully repeating a study serves as a confirmation of its results, and increases the likelihood that the studied hypothesis is correct [350]. Conversely, failing to repeat a study puts the validity of its results into question, in particular when this failure is due to flawed methods. The challenges in reproducing past work has given rise to a perceived 'replication crisis' [248], although this notion is also being challenged [173].

The ability to reproduce studies hinges on the availability and quality of (descriptions of) the data sets, methods and tools used. One set of high-level guiding principles are the FAIR principles [520]: artifacts should be findable, accessible, interoperable, and reusable. Within our field, efforts to support *scientific reproducibility* focus on sharing data sets and tools to allow for repeating studies and building upon prior work. Benzel [81] describes how associations such as ACM [61] and USENIX [482] have an artifact evaluation process where papers can receive badges based on the extent to which artifacts are available, functional, and able to be used for reproducing results. However, these badges may give a false sense of research validity, as the fact that, e.g., methods are reproducible does not mean that they are appropriate or complete [383]. Balenson et al. [74] introduced SEARCCH, an online catalog supporting better discovery of security research artifacts. Hamm et al. [219] found that security papers with user studies generally publish their questionnaires or interview guides, but not the actual participant data that was used in the analysis. More broadly in systems research, Frachtenberg [183] found that the availability of artifacts quickly decays over time. In web measurement research, Demir et al. [141] evaluated recent work on 18 criteria that enable replicability and reproducibility, finding that they often fail to meet these criteria and omit crucial information that would allow reproduction. Hantke et al. [220] evaluated how web archives can be a viable source of historical, reproducible data for web security measurements.

The Internet measurement community has recently made reproducibility a topic of community debate and academic work. Reproducibility was the focus of a workshop at the 2017 SIGCOMM conference. Based on this workshop, Bajpai et al. [72], Saucez and Iannone [423] and Scheitle et al. [428] identified challenges for reproducibility, including ambiguous definitions, unavailability of authors or artifacts, and a lack of incentives. They formulate recommendations to improve reproducibility such as artifact review and badges. Notably, the IMC conference has not implemented such a review and badging process, unlike the security community. Bonaventure [93] and Flittner et al. [181] surveyed authors at computer networking conferences on the composition and availability of paper artifacts. Among their findings, they discuss obstacles such as insufficient descriptions of software and data sets, incomplete tools or broken links, and the influence of research cultures on the type of tools and data sets used, which impacts artifact availability. In 2018, reproducibility was the subject of a Dagstuhl Seminar [70], which resulted in a set of recommendations and best practices for documenting the research process to allow for reproduction [71]. Zilberman and Moore [539] describe experiences with and recommendations for the artifact evaluation process at networking conferences. IMC 2019 featured a 'reproducibility track' [104], inviting short papers replicating prior work, but these were only presented as posters, i.e., not featured at the main conference track.

Evaluation The primary way of *evaluating* research is through the *peer review process*, where fellow scientists judge the quality of a research paper, such as the soundness of its methods or the originality of its findings, and decide whether it is acceptable for formal publication. This process is meant to maintain the integrity of science [436]. However, as peer review remains a human endeavor, concerns prevail about subjectivity in the review process leading to subpar papers with fundamental flaws being published while papers that advance the state of the art are rejected. Ultimately, this could lead to spreading false scientific beliefs and hindering scientific progress, respectively.

In 2022, Soneji et al. [453] studied the peer review process in computer security through interviews with PC[1] members for top-tier conferences. Among their key findings, they found that reviewers did not share common evaluation metrics. Only novelty was a metric considered by most reviewers, although they acknowledged that this was a subjective metric. In contrast, 'red flags' that give reason to reject a paper are more diverse and concrete. This suggests that reviewers may have a mindset of looking for reasons to reject rather than accept papers. While reviewers felt the responsibility to provide high-quality reviews, high workloads, a lack of accountability, and a PC that has insufficient expertise or experience to review a paper run counter to this goal. These yield a risk of subjective reviews and contributes to a sense of 'randomness' as to whether a paper is deemed scientifically worthy. One 'countermovement' to the focus on novelty is the increased appreciation for Systemization of Knowledge papers, which evaluate and systematize existing knowledge on a specific research topic [82]. Specifically for usable security and privacy, Ortloff et al. [370] surveyed reviewers on their criteria for qualitative studies.

[1]The collective of reviewers for one scientific conference is also known as the 'program committee' or PC.

Overall, the reviewers expected detailed methods descriptions and the use of some method for reaching agreement among coders. There was more disagreement on acceptable task division and agreement levels across coders.

The top-tier security conferences have recently moved to a more journal-style model, with multiple submission deadlines and the possibility of revisions. As one possible word of encouragement, Vardi [492] posits that the time and workload pressure brought about by the preference in computer science for conferences over journals reduces review quality. The trend ostensibly started with IEEE S&P adopting rolling deadlines in 2018 [102]. Interestingly, IEEE S&P has since started to backtrack, scrapping revisions for its 2024 edition, due to a concern that papers were no longer being immediately accepted, but instead (unnecessarily) put through a revision process to cater to reviewer interests [112]. The top conferences also start to give more attention to encouraging good reviewing practices, including adding public meta-reviews, avoiding re-reviewing by the same reviewers of a resubmitted paper [112] or recognition through awards. Frachtenberg and Koster [184] surveyed authors of papers at systems conferences, including top security venues. Among their findings, they conclude that authors find review rebuttals and longer reviews very valuable. Sion [445] discusses the shortcomings of the peer review process for computer science conferences from his viewpoint as a PC chair, and proposes to request reviewers to rate more papers more favorably to then increase agreement on whether a paper should be accepted. Lee [294] laments a "toxic culture of rejection" with computer science conferences chasing low acceptance rates, with rejections of otherwise high-quality papers on the basis of lack of novelty or obviousness causing "detrimental effects" to the community.

The Internet measurement and computer networking community has had a longer (academic) experience and experimentation regarding the peer review process. In 2005, Feldmann reported on her experience organizing a 'shadow PC' (also called 'student PC') for SIGCOMM 2005 [177], a parallel PC of mostly junior researchers that runs similarly to a real PC but does not actually decide on the papers that are accepted to the conference. The goal is to give novice researchers an opportunity to experience the review process first hand. Among the findings, Feldmann discussed the differences in paper decisions between the actual and shadow PC, observed a more varied review depth and breadth for the shadow PC, and noted that the experience was well received. The concept of a shadow PC also made it to some editions of security conferences, e.g., USENIX Secuirty in 2014 and 2015, and IEEE S&P from 2016 to 2021. In 2008, Mogul and Anderson [340] summarized prior and future work on best practices for organizing the conference review process. Schulzrinne [431] opines that double-blind reviewing, where authors are anonymous to reviewers, improves perceived fairness but must be implemented judiciously to account for its unintended side effects and limitations such as properly addressing submitted papers that build upon prior publications. Beverly and Allman introspectively measured the IMC 2010 review process [87], with the goal of improving transparency and the process itself. They focused in particular on whether review biases can be measured empirically. The 2011 through 2013 editions of the IMC conference published (meta-)reviews openly, but a community survey led to this practice being discontinued as there were no apparent benefits [28]. Keshav [261] commented

on the "spirit of harsh criticism" that led to an attitude in measurement conferences and the computer science field at large of finding reasons to reject rather than accept a paper. Mogul [339] provided advice on how to reduce 'hypercriticality' and negativity in the reviewing process.

Incentives *Rewarding* research involves evaluating the quality, value, and impact of research, and providing the right incentives and support for research, including appropriate funding. Based on the conference acceptance rate and community input, several conference rankings are used as indicators for quality, both specific to security and privacy venues [210, 537] and for all of computer science [126, 260], with the 'top-tier' conferences being the most attractive and easiest to identify [282]. There is a connection to the peer review process, as the restrictiveness of selecting papers there leads to a division of conferences into tiers of prestige and selectivity. For example, Ortloff et al. [370] commented that replication of qualitative usable security and privacy studies is worthwhile for improving insights, but that such papers may struggle to be accepted to highly valued conferences, therefore disincentivizing researchers from taking the risk of doing such "underappreciated" replication work given a "publish or perish" culture. Publication counts at the most reputable conferences are also used to compile rankings of researchers (e.g., Balzarotti's 'System Security Circus' [75]) and/or institutions (e.g., CSRankings [84]), next to survey-based approaches for the latter [491]. Such rankings are not considered reliable or useful by all, with criticism ranging from questionable methods for survey-based rankings [85, 491] to biases towards established, US-based, 'traditional' institutions and conferences [213]. More fundamentally, such rankings and the data they are based on may say very little about actual quality or other aspects that are harder to measure.

Next to assigning value to a research work based on where it is published, *citations* by other papers are usually used to measure the subsequent impact on the academic field. Rieck [403] maintains a list of highly cited security papers, again only at 'tier 1' and 'tier 2' conferences. Wendzel et al. [514] measured potential factors influencing the citation count of information security papers, using bibliometrics to draw conclusions that, a.o., papers with longer abstracts and more references are cited more often, as well as journal papers, although they also suggest this may be due to a higher number of low-tier conferences with many papers with few citations skewing the data. Vrhovec et al. [502] expanded this analysis, with a contrasting finding that top conference papers are cited more often than journal papers, and described how paper title lengths and references may impact citation counts. While these findings may be statistically validated, there is however no proposed theory that would clearly explain these trends. Overall, the creators of these rankings and counts are often quick to stress that they are merely informal metrics [75, 210, 403, 537] and "are insufficient to characterize all aspects contributing to the relevance of scientific work" [403]. Citations, venue reputation, and quality may also have little relation to each other [139]. 'Altmetrics' are designed to measure research impact online beyond only citations, comprising metrics such as read counts, social media mentions, or media coverage [22]. However, these may not (yet) be a viable alternative [128].

Particular attention also goes to *incentivizing good scientific practices* beyond pure publications. For example, for reproducibility, Collberg and Proebsting [123] proposed additional research funding tied to enforceable 'sharing contracts' in systems research. As another potential incentive, Zheng et al. [536] found that security papers that create and share data sets are likely to be cited more often. Frachtenberg [183] found that systems papers with shared artifacts were cited around 75% more often than those without. On the front of evaluation and peer reviews, Crowcroft et al. [130] proposed mechanisms to incentivize authors, reviewers, and the community to submit higher-quality papers and reviews as well as reward reviewing, and therefore improve the review process. Longstaff et al. [308] found that the time pressure to publish ('breakthrough') results reduces the quality of research experiments and a worse application of a scientific approach. They suggested funding agencies could incentivize security research work that is more based in the scientific method.

Conclusion From our overview of meta-research in web security, we can see that gradually more work is being published on this topic, with varying emphasis on the different categories of meta-research. A strong focus is put on improving methods, especially from an observation that significant pitfalls may be prevalent due to a lack of awareness or critical study. Conducting analyses of state-of-the-art methods can therefore help researchers to select the most appropriate methods and data sets for their study. However, enforcing the use of such methods appears to be a task left for the peer review process, where (individual) reviewers are expected to be aware of current best practices and require that submitted work applies them. Other aspects are enforced or encouraged more strictly or explicitly: for example, ethical considerations become a requirement, artifact evaluation supports reproducibility, and review processes incorporate revisions and public reviews. These fit a trend towards aspiring higher scientific rigor and objectivity.

Improving the soundness and validity of research should be a collective community effort, and there should be venues where the processes and practices that form research can be discussed. For example, in computer security, the *Cyber Security Experimentation and Test* (CSET) and *Learning from Authoritative Security Experiment Results* (LASER) workshops are of interest. Simultaneously, the community can learn from the experiences of other research communities, as was illustrated throughout with examples from the Internet measurement community – observe for example how a top Internet measurement conference stopped publishing meta-reviews in 2013 due to an apparent lack of benefits, yet a top-tier security conference introduced them ten years later. Through this iterative process of reflecting about the way in which security research is conducted, implementing improvements, and evaluating how effective they are – i.e., applying the scientific process to study our research –, this research can become more reliable and trustworthy of itself, and by proxy, help to ensure that this research proves to be beneficial for improving the state of security.

1.2 Dissertation outline

In this dissertation, we cover three case studies selected from the research done during the doctoral program, divided in two parts that each have an additional introduction into the topic they treat. These case studies were chosen for their affinity with meta-research, as they both contribute to critically analyzing current web security data sets or systems that are commonly held to be reliable, while we also propose improved methods as well as discuss data set considerations. Within the previously introduced framework by Ioannidis et al. [249], our contributions lie on the fronts of methods and reproducibility. In general, our goal is to highlight the importance of, and the challenges that exist for sound data sets and methods, which allows us to better understand and even improve how we approach web security research in a valid and reliable manner.

In Part I (Chapters 2 to 4), we present our work analyzing and improving rankings of the most popular websites or domain names on the Internet. These rankings form an important data source for many web security, privacy, and Internet measurement studies. We show how previously commonly used rankings hold potentially undesirable properties that endanger the soundness, validity, and reproducibility of research. We also propose the novel Tranco ranking that improves upon these properties. Tranco combines existing rankings transparently, aggregates across a 30-day period by default to improve long-term stability, and the resulting ranking is made available in a reproducible manner. We confirm with a long-term evaluation that Tranco better matches the properties desirable for research usage. More generally, the Tranco case study addresses how there is a need for open and transparent data sets to enable more valid and sound research practices and reduce biases due to data set errors or false assumptions on their properties.

In Part II (Chapters 5 to 7), we present two case studies on large-scale automated decision-making systems. These are seen as essential tools for processing security-related decisions at scale, and are commonly deployed to handle critical security tasks. However, even error rates that are low on a relative scale can translate into a high error count in absolute terms, which can cause significant harm. In our first case study, we develop a hybrid approach to resolve collisions between benign and malicious domains generated and used by the Avalanche botnet. As erroneous law enforcement decisions would result in unjustified website takedowns and the risk of the botnet reemerging respectively, we involve a human investigator for those domains where an automated model is least certain. In our second case study, we audit Facebook's enforcement of its self-developed policy on political ads. We find that even simple rules for detecting violating ads are not implemented, while many benign ads are falsely taken down, suggesting Facebook's enforcement is imprecise. For both case studies, data set availability forms a crucial component to the ability to effectively conduct the research and develop improved security solutions. For Avalanche, we highlight how we needed to adapt our methods to account for missing data, through an ensemble model, as well as the implications of the unavailability of, a.o., WHOIS data, which is an important factor in our model's performance. For Facebook, we highlight how we needed to develop a novel data collection approach to compensate for the insufficient API-based access that Facebook provided, as it was the only way to

obtain the data necessary to comprehensively audit Facebook's enforcement systems across all possible error types.

Part III (Chapter 8) concludes the dissertation with closing remarks on enablers and challenges for research in our field, focusing on the importance of data sets to analyzing security issues and ecosystems, and developing improved security solutions. We also provide an outlook on future research topics that explore gaps in the current state of the art, in domain rankings and large-scale web measurements in general, as well as automated decision-making systems, with a particular interest for meta-research.

Part I

Analyzing and improving domain rankings

2
Prologue

An adapted version of this prologue was published as a short paper for the 2022 ACSAC Cybersecurity Artifacts Competition and Impact Award.

2.1 Domain rankings

In the next two chapters, we study one prominent data source in web security and privacy as well as Internet measurement research in depth: rankings of the most popular or "top" domain names.[1] These rankings serve multiple purposes in research. First, they allow to select a meaningful sample of domain names for measuring the prevalence of a certain phenomenon or vulnerability, or evaluate a novel technique, attack, or defense. With an estimated 350 million registered domain names [494], it is often *infeasible* to study all domains on the Internet, in terms of the resources required, in particular when collecting data by visiting web pages through an instrumented browser. For example, in one recent study, crawling 50 pages on 100,000 websites using Chromium took 8 days across 60 parallel crawlers [351]. Many of these websites and domain names are also relatively *uninteresting*: they may be special-purpose domains that are unlikely to be visited by the public at large. Selecting only relatively popular domains provides a more representative view on what end users may encounter while they browse the web. Second, the relative ranking of domains also allows to contextualize research findings, indicating whether the popularity of a domain correlates with the occurrence of a certain phenomenon. Third, popular domains are sometimes considered as a source of benign domains, e.g., in a classifier of benign and malicious domains, although this is not a necessary or given property of these rankings, as we show in Section 3.3.5. These rankings form an essential

[1] These rankings are sometimes also called '*(web)sites* rankings' (or 'lists'), referring to the common case where a ranking only contains domains that host a web service. The terms 'sites' and 'domains' are sometimes used interchangeably or confused. We will usually examine domain rankings, of which we consider website rankings to be a subset.

part of many web and Internet research studies: in our 2018 literature survey (Section 3.4.1), we found 133 papers across four years in top-tier security conferences to use at least one ranking, and over 400 studies have used our Tranco ranking (Chapter 3) at the time of writing.

Analyzing domain rankings Historically, major commercial players published domain rankings. At the time of the first work presented in this chapter, these were Alexa (since 2008), Quantcast (since 2007), and later Majestic (since 2012) in the SEO and marketing space, all publishing rankings containing only websites. (Cisco) Umbrella also published a ranking (since 2016) as a large DNS resolver, which included any type of domain name, i.e., also infrastructural domains. Despite the importance of these commercial rankings to research and decades of use, the research community only scrutinized them and became fully aware of the issues surrounding them from 2018 onward. Two contemporaneous works studied these rankings in depth: the *"A long way to the top"* paper by Scheitle et al. [427], and our Tranco paper (Chapter 3). They reverse engineered the proprietary and opaque methods used to construct the four commercial rankings, surveyed their usage in research usage, and measured characteristics such as stability and similarity, and quantified the potential impact on Internet measurement and security research respectively. These works both found low agreement, high volatility in some lists, and therefore a potentially large impact on research results. In our Tranco work, we also showed that all four rankings could be manipulated to insert any domain, which has, a.o., implications on their use in whitelists.

After the publication of these works, Rweyemamu et al. studied specific aspects such as weekly patterns and domain clusters in more detail [416], and refined the manipulation attacks for Alexa and Umbrella [417]. We evaluated our Tranco ranking on the same properties as the work by Scheitle et al. and our work on the commercial rankings (Chapter 4), to find that the proposed default parameters form a good common ground for use in research. Alby and Jäschke [32] expand the comparison of top lists to other data sets such as Wikipedia, Common Crawl, and search engine results, in terms of overlap and agreement on popularity. They also find low overlap, with hosts popular in one data set missing from others. They conclude with a recommendation for random sampling among Common Crawl hosts. Ruth et al. [415] compared top lists to Cloudflare traffic data, coming closest to developing a 'ground truth' for website popularity. They concluded that the Chrome User Experience Report (CrUX) most accurately represents the more popular websites, albeit as an unordered set, although agreement remains relatively low. Tranco inherits its accuracy and biases from its component lists, ending up with worse accuracy than CrUX. The CrUX list comes with its own limitations that make it less usable for certain research purposes: it does not cover any infrastructural domains (i.e., beyond websites), is only updated monthly, and is primarily only accessible using an SQL query to its BigQuery data set. Nevertheless, it is a valuable data set, and we consider adding it to Tranco (Section 8.2.1). Industry companies have also published independent analyses of Tranco. Infoblox compared their (proprietary and not publicly available) InfoRanks ranking [471] to existing rankings, including Tranco, on properties such as similarity

and benignness. DeepSee analyzed the influence of individual component rankings, and in particular Alexa, on the final ranks in Tranco [346].

Improving domain rankings In addition to analyzing and describing the issues with existing rankings, two academic initiatives design and publish new research-oriented domain rankings, with a strong focus on improving upon properties important for research, such as reproducibility and transparency. In 2019, we made the Tranco list available (Chapter 3). The construction method is public, and by default consists of aggregating multiple existing rankings across a 30-day period to improve stability. Daily generated and archived rankings are published online.[2] In 2022, Xie et al. [523] developed the SecRank ranking method for passive DNS traffic. The global ranking is based on weighted voting across the domain preferences of individual IP addresses. In the evaluation, they apply their method to passive DNS data from 114 DNS (a Chinese public DNS resolver), and publish this list.[3] Since IP-level voting is crucial to their method, the DNS request data from individual users are required, which may be undesirable from the privacy perspective [518]. In industry, 2022 yielded two new public passive DNS-based rankings from Cloudflare ('Radar') [320] and Webshrinker ('DNSFilter') [309].

The concept of ranking has also been extended to other Internet properties. Naab et al. [352] developed and published prefix top lists[4] that derive a ranking of network prefixes from existing domain rankings, intended for measurements of Internet infrastructure such as CDNs or core routers. Aqeel et al. [58] developed and published Hispar,[5] a ranking that includes internal pages, i.e., pages beyond the landing pages that are typically measured when visiting the websites in a domain ranking. They discover these internal pages through search engine results. The ranking is intended for increasing the per-website coverage of a measurement, and improve the validity of findings for a website. Unfortunately, both initiatives appear to have stopped publishing new rankings. Marquardt and Schmidt [319] propose to use Certificate Transparency logs to generate a domain list that is larger and more diverse, by covering as many hosts as possible, i.e., not only popular domains.

2.2 Usage and impact of Tranco

Since its original publication in 2019 (Chapter 3), the selection of component rankings in the Tranco ranking has been updated, due to the disappearance or emergence of lists. The Quantcast ranking was silently discontinued and subsequently removed in April 2020. The Alexa ranking is also no longer updated since February 2023, and is slated to be removed from Tranco as well. DomainTools allowed their passive DNS-based 'Farsight' ranking [195] to be integrated into the default Tranco list, but it is otherwise not public.

[2] https://tranco-list.eu/
[3] https://secrank.cn/topdomain
[4] https://prefixtoplists.net.in.tum.de/
[5] https://hispar.cs.duke.edu/

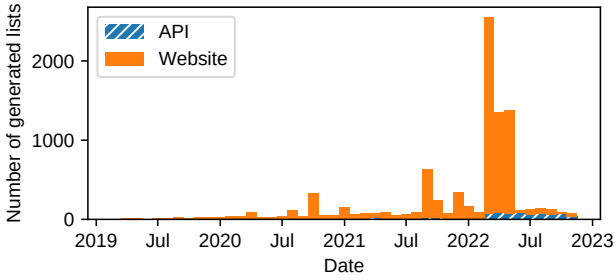


Figure 2.1: Number of customized lists generated over time, through the website or API.

Within their threat intelligence product Iris, they also incorporate a (private) ranking that is aggregated using the same method as Tranco [195]. In Section 8.2.1, we discuss future directions for Tranco, including the inclusion of additional rankings, in more detail. In the rest of this section, we give an overview of how Tranco has been used within and outside the research community.

Usage statistics The online service for accessing the daily standard Tranco ranking and generating customized lists is publicly and freely available. Since the Tranco ranking was made public in February 2019 until 21 November 2022, 1,374 daily lists were published on this service. In addition, 497 distinct users generated 9,268 customized lists, with a noticeable uptick for the latter in 2022 (Figure 2.1). The API is also increasingly popular for generating lists, having been used for 712 lists already since early 2022. Filters are used in moderation, with 858 lists selecting probable websites using the Chrome User Experience Report, 399 lists including or excluding certain TLDs, 346 lists removing known malicious websites using Google Safe Browsing, and 262 lists filtering organizations.

According to a one-week log from September 2022 of the AWS S3 bucket serving the daily list, around 25 unique endpoints request the daily list per hour, with a noticeable spike to around 70 endpoints when the new daily list is released at midnight (Figure 2.2). Based on web analytics data available to us, the Tranco website receives at least 700 daily unique visitors, and over 1,000 monthly clickthroughs from Google search results. These last metrics ignore those who opt out of analytics tracking, as well as those who directly access list ZIP files, in particular the daily list, which is available at a stable URL that does not require passing through the Tranco website.

The GitHub repository with our open-source implementation has been forked 12 times and starred 74 times, although this underrepresents the impact, as not many people will need to consult the list generation code but instead use the end product that is available on our website. More telling is the usage of the Tranco Python package, which has been downloaded over 121,000 times [51], with a noticeable increase in fall 2021 (Figure 2.3). This shows that our open dataset is integrated into a variety of software projects.

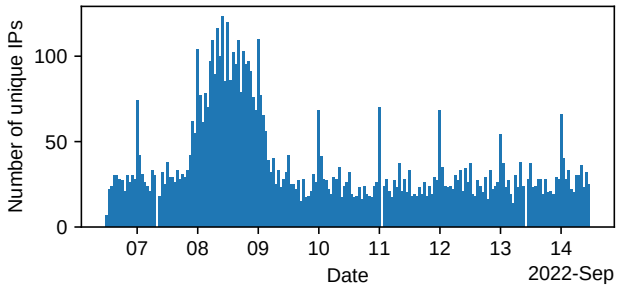


Figure 2.2: Number of unique IPs downloading the daily list, grouped by hour.

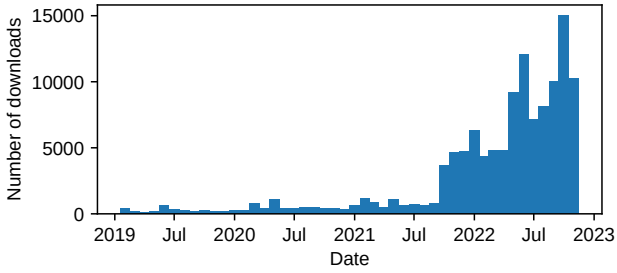


Figure 2.3: Number of monthly downloads of the Tranco Python package.

Academic impact Tranco has demonstrated a large impact on web and Internet measurement research in both academia and beyond. The most readily measurable impact is the number of citations in academic research papers, currently amounting to over 400 in 4 years, according to Google Scholar[6], therefore being one of the top 10 cited security papers from 2019[7]. Tranco has been used in publications at the four top-tier security conferences and their co-located workshops, other security and privacy conferences, as well as the prominent Web and Internet measurement conferences (Table 2.1). Tranco has also found its way into other research domains, e.g., being used in the process of creating neural models for information extraction from websites [419].

Tranco has arguably become the *de facto* standard top list in the research field. Demir et al. [141] highlighted Tranco as a "work[] that aim[s] to provide best practices" and that "ha[s] a positive impact on our community." Anecdotally, we know reviewers are aware of this position. We have received comments from reviewers either suggesting we use Tranco instead of Alexa or commending us for using Tranco, as well as pointing out the reproducibility harm when omitting the exact Tranco list used in a study. Generally,

[6] https://scholar.google.be/scholar?oi=bibs&cites=1499698348405075976,10234769677796230547,17897712023882147302

[7] https://www.mlsec.org/topnotch/sec_2010s.html

Table 2.1: Distribution of papers citing Tranco across venue types and years, across the 250 papers for which we were able to retrieve bibliographic data. * Until August 2022.

	2019	2020	2021	2022*	Total
Top-tier security conferences	1	7	15	12	35
co-located workshops	6	3	6	4	19
Other security conferences	6	9	12	11	38
Privacy conferences	0	0	9	3	12
Web & Internet measurement	9	14	13	10	46
Other conferences	6	9	12	8	35
Journals	2	6	9	11	28
Unknown	3	7	11	16	37
Total	33	55	87	75	250

we see that papers mention the Tranco project and the exact list ID well, providing an immediate benefit to their reproducibility and validity.

Industry and media impact The Tranco list is used by prominent industry players, showing its wide applicability. Tranco is a contributing data partner to projects by the Electronic Frontier Foundation (for their 'Privacy Badger' extension) [331], the Internet Society's Pulse project (measuring the health, availability and evolution of the Internet) [246], the Brave browser (filtering search results) [98], ScamAdviser (as a popularity indicator),[8] URLhaus by abuse.ch (sharing malware URLs) [479], BuiltWith [99], and W3Techs [196] (the latter two measuring web technology usage). Tranco has been used for measurements by Mozilla [470], Cloudflare [495], F5 Labs [507], Palo Alto Networks' Unit 42 [303], Avast [95], and ICANN [228, 229]. Tranco is available as a filter list in the threat intelligence platforms from MISP [337], SEKOIA [477], Intel Owl [77], and ThreatConnect [199].

On the media side, Tranco was used by journalists from the Norwegian public broadcaster NRK [211] and The Markup [323, 421], and was mentioned in news articles from The Verge [496] and SecurityWeek [272] that reported on the demise of the Alexa ranking. Prominent security researchers Scott Helme and Troy Hunt use Tranco for their scans of the state of security on the web [225] and the Why No HTTPS? project [237], respectively.

2.3 Overview

In Chapter 3, we first analyze the composition of four commercial top domain rankings, finding that they may have inherent properties that are undesirable for research. We

[8] see, e.g., `https://www.scamadviser.com/check-website/scamadviser.com`

then empirically validate that these rankings are vulnerable to manipulation. Finally, we provide the Tranco ranking, a research-oriented ranking that is publicly available and emphasizes reproducibility and reliability. In Chapter 4, we analyze the Tranco ranking longitudinally across the properties studied in our previous work. We find that the default parameters of Tranco result in a stable, robust and comprehensive ranking, and provide recommendations to researchers for using Tranco.

3

Tranco: A Research-Oriented Top Sites Ranking Hardened Against Manipulation

This chapter is based on the homonymous paper published in the proceedings of the 26th Annual Network and Distributed System Security Symposium (NDSS 2019) [291]. This work was co-authored with Tom Van Goethem, Samaneh Tajalizadehkhoob, Maciej Korczyński, and Wouter Joosen.

In order to evaluate the prevalence of security and privacy practices on a representative sample of the Web, researchers rely on website popularity rankings such as the Alexa list. While the validity and representativeness of these rankings are rarely questioned, our findings show the contrary: we show for four main rankings how their inherent properties (similarity, stability, representativeness, responsiveness and benignness) affect their composition and therefore potentially skew the conclusions made in studies. Moreover, we find that it is trivial for an adversary to manipulate the composition of these lists. We are the first to empirically validate that the ranks of domains in each of the lists are easily altered, in the case of Alexa through as little as a single HTTP request. This allows adversaries to manipulate rankings on a large scale and insert malicious domains into whitelists or bend the outcome of research studies to their will. To overcome the limitations of such rankings, we propose improvements to reduce the fluctuations in list composition and guarantee better defenses against manipulation. To allow the research community to work with reliable and reproducible rankings, we provide Tranco, an improved ranking that we offer through an online service available at `https://tranco-list.eu`.

3.1 Introduction

Researchers and security analysts frequently study a selection of popular sites, such as for measuring the prevalence of security issues or as an evaluation set of available and often used domain names, as these are purported to reflect real-world usage. The most well known and widely used list in research studies is that of Alexa, with researchers' reliance on this commercial list being accentuated by their concern when it was momentarily taken offline in November 2016 [39]. However, several companies provide alternative rankings based on Internet usage data collected through various channels [354]: a panel of users whose visits are logged, tracking code placed on websites and traffic captured by intermediaries such as ISPs.

We found that 133 top-tier studies over the past four years based their experiments and conclusions on the data from these rankings. Their validity and by extension that of the research that relies on them, should however be questioned: the methods behind the rankings are not fully disclosed, and commercial interests may prevail in their composition. Moreover, the providers only have access to a limited userbase that may be skewed towards e.g. certain user groups or geographic regions. Even though most providers declare that the data is processed to remove such statistical biases, the lack of exact details makes it impossible for researchers to assess the potential impact of these lists on their results and conclusions.

In this paper, we show that the four main popularity rankings (Alexa, Cisco Umbrella, Majestic and Quantcast) exhibit significant problems for usage in research. The rankings hardly agree on the popularity of any domain, and the Umbrella and especially the Alexa lists see a significant turnover even on consecutive days; for Alexa, this is the result of an unannounced and previously unknown change in averaging approach. All lists include non-representative and even malicious sites, which is especially dangerous considering the widespread use of these rankings as whitelists. Overall, these flaws can cause the choice for a particular ranking to severely skew measurements of vulnerabilities or secure practices.

Moreover, we are the first to empirically prove that pitfalls in these rankings leave them vulnerable to one of our newly introduced manipulation techniques. These techniques have a surprisingly low cost, starting from a single HTTP request for Alexa, and can therefore be used to affect the rank of thousands of domains at once on a substantial level: we estimate that the top 10 000 can easily be reached. The incentives of adversaries to alter the composition of these lists, both for single domains due to the practice of whitelisting popular domains, and on a larger scale to influence research and its impact outside academia, make this manipulation particularly valuable.

Finally, there is still a need for researchers to study popular domains, so they would therefore benefit from a list that avoids biases in its inherent properties and is more resilient to manipulation, and that is easily retrieved for future reference. To this extent, we propose improvements to current rankings in terms of stability over time, representativeness and hardening against manipulation. We create Tranco, a new

ranking that is made available and archived through an accompanying online service at `https://tranco-list.eu`, in order to enhance the reproducibility of studies that rely on them. The community can therefore continue to study the security of popular domains while ensuring valid and verifiable research.

In summary, we make the following contributions:

- We describe how the main rankings can negatively affect security research, e.g. half of the Alexa list changes every day and the Umbrella list only has 49% real sites, as well as security implementations, e.g. the Majestic list contains 2 162 malicious domains despite being used as a whitelist.
- We classify how 133 recent security studies rely on these rankings, in particular Alexa, and show how adversaries could exploit the rankings to bias these studies.
- We show that for each list there exists at least one technique to manipulate it on a large scale, as e.g. only one HTTP request suffices to enter the widely used Alexa top million. We empirically validate that reaching a rank as good as 28 798 is easily achieved.
- Motivated by the discovered limitations of the widely-used lists, we propose Tranco, an alternative list that is more appropriate for research, as it varies only by 0.6% daily and requires at least the quadrupled manipulation effort to achieve the same rank as in existing lists.

3.2 Methodology of top websites rankings

Multiple commercial providers publish rankings of popular domains that they compose using a variety of methods. For Alexa, Cisco Umbrella, Majestic and Quantcast, the four lists that are available for free in an easily parsed format and that are regularly updated, we discuss what is known on how they obtain their data, what metric they use to rank domains and which potential biases or shortcomings are present. We base our discussion mainly on the documentation available from these providers; many components of their rankings are proprietary and could therefore not be included.

We do not consider any lists that require payment, such as SimilarWeb[1], as their cost (especially for longitudinal studies) and potential usage restrictions make them less likely to be used in a research context. We also disregard lists that would require scraping, such as Netcraft[2], as these do not carry the same consent of their provider implied by making the list available in a machine-readable format. Finally, Statvoo's list[3] seemingly meets our criteria. However, we found it to be a copy of Alexa's list of November 23, 2016, having never been updated since; we therefore do not consider it in our analysis.

[1] `https://www.similarweb.com/top-websites`
[2] https://toolbar.netcraft.com/stats/topsites
[3] `https://statvoo.com/dl/top-1million-sites.csv.zip`

3.2.1 Alexa

Alexa, a subsidiary of Amazon, publishes a daily updated list[4] consisting of one million websites since December 2008 [35]. Usually only pay-level domains[5] are ranked, except for subdomains of certain sites that provide 'personal home pages or blogs' [36] (e.g. `tmall.com`, `wordpress.com`). In November 2016, Alexa briefly took down the free CSV file with the list [39]. The file has since been available again [38] and is still updated daily; however, it is no longer linked to from Alexa's main website, instead referring users to the paid 'Alexa Top Sites' service on Amazon Web Services [48].

The ranks calculated by Alexa are based on traffic data from a "global data panel", with domains being ranked on a proprietary measure of unique visitors and page views, where one visitor can have at most one page view count towards the page views of a URL [528]. Alexa states that it applies "data normalization" to account for biases in their user panel [36].

The panel is claimed to consist of millions of users, who have installed one of "many different" browser extensions that include Alexa's measurement code [37]. However, through a crawl of all available extensions for Google Chrome and Firefox, we found only Alexa's own extension ("Alexa Traffic Rank") to report traffic data. Moreover, this extension is only available for the desktop version of these two browsers. Chrome's extension is reported to have around 570 000 users [40]; no user statistics are known for Firefox, but extrapolation based on browser usage suggests at most one million users for two extensions, far less than Alexa's claim.

In addition, sites can install an 'Alexa Certify' tracking script that collects traffic data for all visitors; the rank can then be based on these actual traffic counts instead of on estimates from the extension [36]. This service is estimated to be used by 1.06% of the top one million and 4% of the top 10 000 [101].

The rank shown in a domain's profile on Alexa's website is based on data over three months, while in 2016 they stated that the downloadable list was based on data over one month [33]. This statement was removed after the brief takedown of this list [34], but the same period was seemingly retained. However, as we derive in Section 3.3.2, since January 30, 2018 the list is based on data for one day; this was confirmed to us by Alexa but was otherwise unannounced.

Alexa's data collection method leads to a focus on sites that are visited in the top-level browsing context of a web browser (i.e. HTTP traffic). They also indicate that ranks worse than 100 000 are not statistically meaningful, and that for these sites small changes in measured traffic may cause large rank changes [36], negatively affecting the stability of the list.

[4] `https://s3.amazonaws.com/alexa-static/top-1m.csv.zip`

[5] A *pay-level domain* (PLD) refers to a domain name that a consumer or business can directly register, and consists of a subdomain of a *public suffix* or *effective top-level domain* (e.g. `.com` but also `.co.uk`).

3.2.2 Cisco Umbrella

Cisco Umbrella publishes a daily updated list[6] consisting of one million entries since December 2016 [236]. Any domain name may be included, with it being ranked on the aggregated traffic counts of itself and all its subdomains.

The ranks calculated by Cisco Umbrella are based on DNS traffic to its two DNS resolvers (marketed as OpenDNS), claimed to amount to over 100 billion daily requests from 65 million users [236]. Domains are ranked on the number of unique IPs issuing DNS queries for them [236]. Not all traffic is said to be used: instead the DNS data is sampled and 'data normalization methodologies' are applied to reduce biases [119], taking the distribution of client IPs into account [310]. Umbrella's data collection method means that non-browser-based traffic is also accounted for. A side-effect is that invalid domains are also included (e.g. internal domains such as `*.ec2.internal` for Amazon EC2 instances, or typos such as `google.conm`).

3.2.3 Majestic

Majestic publishes the daily updated 'Majestic Million' list consisting of one million websites[7] since October 2012 [254]. The list comprises mostly pay-level domains, but includes subdomains for certain very popular sites (e.g. `plus.google.com`, `en.wikipedia.org`).

The ranks calculated by Majestic are based on backlinks to websites, obtained by a crawl of around 450 billion URLs over 120 days, changed from 90 days on April 12, 2018 [315, 316]. Sites are ranked on the number of class C (IPv4 /24) subnets that refer to the site at least once [254]. Majestic's data collection method means only domains linked to from other websites are considered, implying a bias towards browser-based traffic, however without counting actual page visits. Similarly to search engines, the completeness of their data is affected by how their crawler discovers websites.

3.2.4 Quantcast

Quantcast publishes a list[8] of the websites visited the most in the United States since mid 2007 [394]. The size of the list varies daily, but usually was around 520,000 mostly pay-level domains; subdomains reflect sites that publish user content (e.g. `blogspot.com`, `github.io`). The list also includes 'hidden profiles', where sites are ranked but the domain is hidden.

The ranks calculated by Quantcast are based on the number of people visiting a site within the previous month, and comprises 'quantified' sites where Quantcast directly

[6] `https://s3-us-west-1.amazonaws.com/umbrella-static/top-1m.csv.zip`
[7] `http://downloads.majestic.com/majestic_million.csv`
[8] `https://ak.quantcast.com/quantcast-top-sites.zip`

measures traffic through a tracking script as well as sites where Quantcast estimates traffic based on data from 'ISPs and toolbar providers' [442]. These estimates are only calculated for traffic in the United States, with only quantified sites being ranked in other countries; the list of top sites also only considers US traffic. Moreover, while quantified sites see their visit count updated daily, estimated counts are only updated monthly [443], which may inflate the stability of the list. Before November 14, 2018, quantified sites made up around 10% of the full (US) list. However, since then Quantcast seems to have stopped ranking almost any estimated domains, therefore reducing the list size to around 40 000.

3.3 Quantitative comparison

Ideally, the domain rankings would perfectly reflect the popularity of websites, free from any biases. However, the providers of domain rankings do not have access to complete Internet usage data and use a variety of largely undisclosed data collection and processing methods to determine the metric on which they rank websites. This may lead to differences between the lists and potential 'hidden' factors influencing the rankings: the choice of list can then critically affect e.g. studies that measure the prevalence of security practices or vulnerabilities. We compare the four main lists over time in order to assess the breadth and impact of these differences.

Certain properties may reflect how accurately Internet usage is measured and may be (more or less) desired when using the lists for security research. We consider five properties in our comparison: 1. *similarity* or the agreement on the set of popular domains, 2. *stability* or the rank changes over time, 3. *representativeness* or the reflection of popularity across the web, 4. *responsiveness* or the availability of the listed websites, and 5. *benignness* or the lack of malicious domains.

To quantitatively assess these properties, we use the lists obtained between January 1 and November 30, 2018, referring to the date when the list would be downloaded; the data used by the provider to compile the list may be older. In addition, we crawled the sites on the four lists as downloaded on May 11, 2018 at 13:00 UTC from a distributed crawler setup of 10 machines with 4 CPU cores and 8 GB RAM in our European university network, using Ubuntu 16.04 with Chromium version 66.0.3359.181 in headless mode.

3.3.1 Similarity

Figure 3.1 shows the average number of sites that the rankings agree upon per day; there is little variance over time. The four lists combined contain around 2.82 million sites, but agree only on around 70 000 sites. Using the rank-biased overlap (RBO) [513], a similarity measure that can be parameterized to give a higher weight to better ranks, we see that the lists of Alexa, Majestic and Quantcast are the most similar to each other. However, even when heavily weighting the top 100, the RBO remains low between 24% and 33%. Umbrella's full list is most dissimilar to the others, with an RBO of between

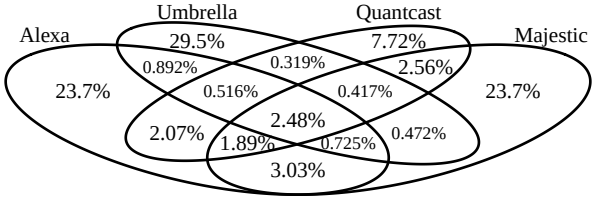


Figure 3.1: The average daily intersections between the lists of the four providers from January 30, 2018 to November 13, 2018.

4.5% and 15.5%. However, this is to be expected as Umbrella includes subdomains: when ranking only pay-level domains, the RBO with the other lists reaches around 30% as well. Finally, Quantcast's removal of non-quantified sites after November 14, 2018 causes a significant drop in RBO to less than 5.5%, with no overlap of the top 10: many very popular domains are not quantified and are therefore now missing from Quantcast's list.

The small overlaps signify that there is no agreement on which sites are the most popular. This means that switching lists yields a significantly different set of domains that can e.g. change how prevalent certain web trackers seem to be [163].

3.3.2 Stability

From the intersections between each provider's lists for two consecutive days, shown in Figure 3.2, we see that Majestic's and Quantcast's lists are the most stable, usually changing at most 1% per day, while for Umbrella's list this climbs to on average 10%. Until January 30, 2018, Alexa's list was almost as stable as Majestic's or Quantcast's. However, since then stability has dropped sharply, with around half of the top million changing every day, due to Alexa's change to a one day average. There exists a trade-off in the desired level of stability: a very stable list provides a reusable set of domains, but may therefore incorrectly represent sites that suddenly gain or lose popularity. A volatile list however may introduce large variations in the results of longitudinal studies.

3.3.3 Representativeness

Sites are mainly distributed over a few top-level domains, with Figure 3.3 showing that 10 TLDs capture more than 73% of every list. The .com TLD is by far the most popular, at almost half of Alexa's and Majestic's list and 71% of Quantcast's list; .net, .org and .ru are used most often by other sites. One notable outlier is the .jobs TLD: while for the other lists it does not figure in the top 10 TLDs, it is the fourth most popular TLD for Quantcast. Most of these sites can be traced to DirectEmployers, with thousands of lowly ranked domains. This serves as an example of one entity controlling a large part of a ranking, potentially giving them a large influence in research results.

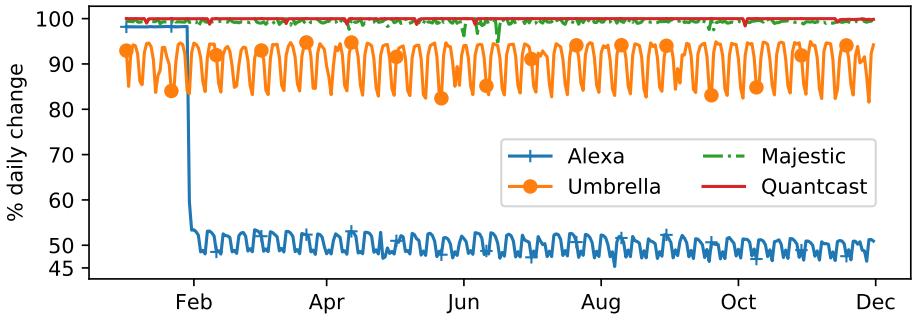


Figure 3.2: The intersection percentage between each provider's lists for two consecutive days.

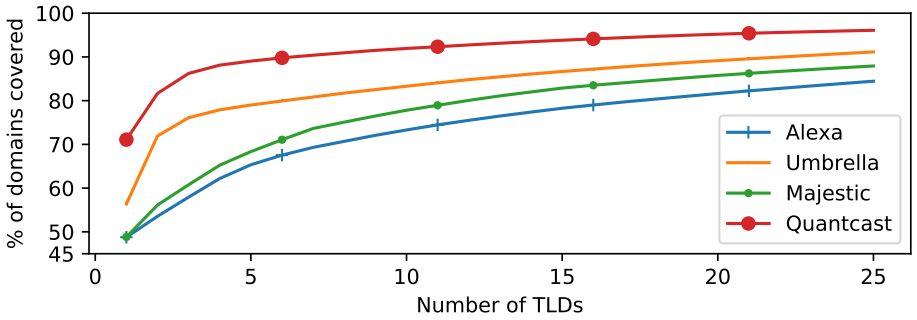


Figure 3.3: The cumulative distribution function of TLD usage across the lists.

We use the autonomous system to determine the entities that host the ranked domains. Google hosts the most websites within the top 10 and 100 sites, at between 15% and 40% except for Quantcast at 4%: for Alexa these are the localized versions, for the other lists these are subdomains. For the full lists, large content delivery networks dominate, with Cloudflare being the top network hosting up to 10% of sites across all lists. This shows that one or a few entities may be predominantly represented in the set of domains used in a study and that therefore care should be taken when considering the wider implications of its results.

3.3.4 Responsiveness

Figure 3.4 shows the HTTP status code reported for the root pages of the domains in the four lists. 5% of Alexa's and Quantcast's list and 11% of Majestic's list could not be reached. For Umbrella, this jumps to 28%; moreover only 49% responded with status code 200, and 30% reported a server error. Most errors were due to name resolution failure, as invalid or unconfigured (sub)domains are not filtered out.

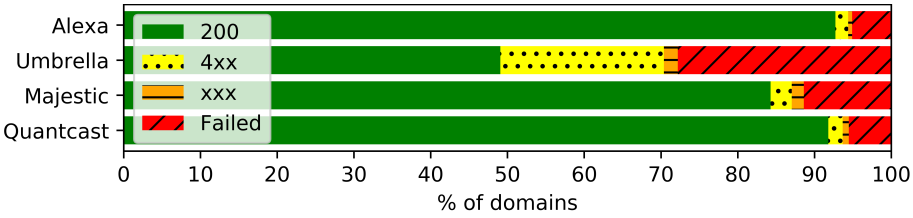


Figure 3.4: The responsiveness and reported HTTP status code across the lists.

Of the reachable sites, 3% for Alexa and Quantcast, 8.7% for Majestic and 26% for Umbrella serve a page smaller than 512 bytes on their root page, based on its download size as reported by the browser instance. As such pages often appear empty to the user or only produce an error, this indicates that they may not contain any useful content, even though they are claimed to be regularly visited by real users. Unavailable sites and those without content do not represent real sites and may therefore skew e.g. averages of third-party script inclusion counts [362], as these sites will be counted as having zero inclusions.

3.3.5 Benignness

Malicious campaigns may target popular domains to extend the reach of their attack, or use a common domain as a point of contact, leading to it being picked up as 'popular'. While it is not the responsibility of ranking providers to remove malicious domains, popular sites are often assumed to be trustworthy, as evidenced by the practice of whitelisting them [193] or, as we show in Section 3.4.1, their usage in security research as the benign test set for classifiers.

Table 3.1 lists the number of domains flagged on May 31, 2018 by Google Safe Browsing, used among others by Chrome and Firefox to automatically warn users when they visit dangerous sites [418]. At 0.22% of its list, Majestic has the most sites that are flagged as potentially harmful (in particular as malware sites), but all lists rank at least some malicious domains. In Alexa's top 10 000, 4 sites are flagged as performing social engineering (e.g. phishing), while 1 site in Majestic's top 10 000 serves unwanted software. The presence of these sites in Alexa's and Quantcast's list is particularly striking, as users would have to actively ignore the browser warning in order to trigger data reporting for Alexa's extension or the tracking scripts.

Given the presence of malicious domains on these lists, the practice of whitelisting popular domains is particularly dangerous. Some security analysis tools whitelist sites on Alexa's list [226, 326]. Moreover, Quad9's DNS-based blocking service whitelists all domains on Majestic's list [193], exposing its users to ranked malicious domains. As Quad9's users expect harmful domains to be blocked, they will be even more under the impression that the site is safe to browse; this makes the manipulation of the list very interesting to attackers.

Table 3.1: Presence of domains in the four rankings on Google's Safe Browsing list on May 31, 2018.

	Malware		Social Engineering			Unwanted software			Potentially harmful application		Total
	100*k*	Full	10*k*	100*k*	Full	10*k*	100*k*	Full	100*k*	Full	
Alexa	32	98	4	85	345	0	15	104	0	0	547
Umbrella	11	326	0	3	393	0	23	232	4	60	1011
Majestic	130	1676	0	23	359	1	9	79	9	48	2162
Quantcast	3	76	0	4	105	0	4	41	0	2	224

3.4 Usage in security research

Whenever security issues are being investigated, researchers may want to evaluate their impact on real-world domains. For these purposes, security studies often use and reference the top sites rankings. The validity and representativeness of these rankings therefore directly affects their results, and any biases may prohibit correct conclusions being made. Moreover, if forged domains could be entered into these lists, an adversary can control research findings in order to advance their own goals and interests.

3.4.1 Survey and classification of list usage

To assess how security studies use these top sites rankings, we surveyed the papers from the main tracks of the four main academic security conferences (CCS, NDSS, S&P, USENIX Security) from 2015 to 2018; we select these venues as they are considered top-tier and cover general security topics. We classify these papers according to four purposes for the lists: *prevalence* if the rankings are used to declare the proportion of sites affected by an issue; *evaluation* if a set of popular domains serves to test an attack or defense, e.g. for evaluating Tor fingerprinting [405]; *whitelist* if the lists are seen as a source of benign websites, e.g. for use in a classifier [499]; *ranking* if the exact ranks of sites are mentioned or used (e.g. to estimate website traffic [163]) or if sites are divided into bins according to their rank.

Alexa is by far the most popular list used in recent security studies, with 133 papers using the list for at least one purpose. Table 3.2 shows the number of papers per category and per subset of the list that was used. The Alexa list is mostly used for measuring the prevalence of issues or as an evaluation set of popular domains. For the former purpose as well as for whitelisting and ranking or binning, the full list is usually used, while for evaluation sets, the subset size varies more widely. Three papers from these conferences also used another ranking, always in tandem with the Alexa list [94, 532, 540].

Table 3.2: Categorization of recent security studies using the Alexa ranking. One study may appear in multiple categories.

	Subset studied								
Purpose	10	100	500	1*k*	10*k*	100*k*	1*M*	Other	Total
Prevalence	1	6	8	9	16	7	32	13	63
Evaluation	7	16	14	10	9	3	14	28	71
Whitelist	0	2	1	4	3	2	11	6	19
Ranking	0	1	3	3	2	4	15	7	28
Total	8	20	18	18	23	9	45	36	133

Most studies lack any comment on when the list was downloaded, when the websites on the lists were visited and what proportion was actually reachable. This hampers reproducibility of these studies, especially given the daily changes in list compositions and ranks.

Two papers commented on the methods of the rankings. Juba et al. [255] mention the rankings being "representative of true traffic numbers in a coarse grained sense". Felt et al. [179] mention the "substantial churn" of Alexa's list and the unavailability of sites, and express caution in characterizing all its sites as popular. However, in general the studies do not question the validity of the rankings, even though they have properties that can significantly affect their conclusions, and as we will show are vulnerable to manipulation.

3.4.2 Influence on security studies

Incentives

Given the increasing interest in cybersecurity within our society, the results of security research have an impact beyond academia. News outlets increasingly report on security vulnerabilities, often mentioning their prevalence or affected high-profile entities [206–208, 485]. Meanwhile, policy-makers and governments rely on these studies to evaluate secure practices and implement appropriate policies [68, 148]; e.g. Mozilla in part decided to delay distrusting Symantec certificates based on a measurement across Umbrella's list [472].

Malicious actors may therefore risk exposure to a wider audience, while their practices may trigger policy changes, yielding them an incentive to directly influence security studies. Invernizzi et al. [247] discovered that blacklists sold on underground markets contain IP addresses of academic institutions as well as security companies and researchers, illustrating that adversaries already actively try to prevent detection by researchers. As we showed, security studies often rely on popularity rankings, so pitfalls

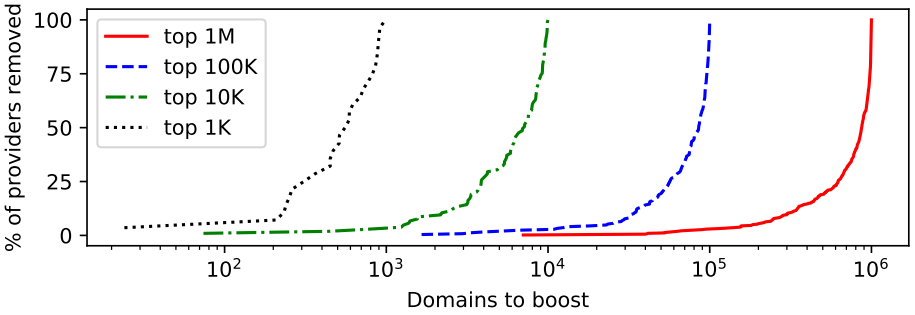


Figure 3.5: The percentage of fingerprinting script providers that would not be detected if a given number of domains were pushed above all fingerprinting domains for different subsets of Alexa's ranking.

in the methods of these rankings that expose them to targeted manipulation open up another opportunity for adversaries to affect security research. The way in which an adversary may want to influence rankings, and therefore the research dependent upon them, varies according to their incentives. They may want to promote domains into the lists, making them be perceived as benign and then execute malicious practices through them. Alternatively, they can promote other domains to hide their own malicious domains from the lists. Finally, they can intelligently combine both techniques to alter comparisons of security properties for websites of different entities.

Case study

The issue of online tracking and fingerprinting has been studied on multiple occasions for Alexa's top one million [163, 281, 299, 300, 362]. Users may want to avoid organizations that perform widespread or invasive tracking, and therefore have an interest in new tracking mechanisms and/or specific trackers being found or named by these studies, e.g. to include them in blocklists. The trackers therefore have an incentive to avoid detection by not figuring among the domains being studied, e.g. by pushing these out of the popularity ranking used to provide the set of investigated domains.

We quantify the effort required to manipulate a ranking and therefore alter findings for the measurements of fingerprinting prevalence by Acar et al. [15] and Englehardt and Narayanan [163] on Alexa's top 100 000 and top one million respectively. These studies published data on which domains included which scripts, including the Alexa rank. We calculate how many domains minimally need to be moved up in order to push out the websites using a particular tracking provider.

Figure 3.5 shows how many fingerprinting providers would fully disappear from the Alexa list if a given number of domains are manipulated. We consider removal for different subsets, as commonly used by the studies that we surveyed in Section 3.4.1. The smallest number of manipulated domains required is 7 032, 1 652, 74 and 24 for the top 1M, 100K,

10K and 1K respectively; 15 providers need less than 100 000 manipulated domains to disappear from the top 1M.

As we will show, the cost of such large-scale manipulation is very low and well within reach of larger providers, especially given the incentive of being able to stealthily continue tracking. Moreover, this is an upper bound needed to remove all instances of a tracker domain from the list; reducing the prevalence of a script requires only hiding the worst-ranked domains. Finally, it is not required to insert new domains: forging a few requests to boost sites already in the list is sufficient, further reducing the cost and even making the manipulation harder to detect.

Englehardt and Narayanan highlighted how "the long tail of fingerprinting scripts are largely unblocked by current privacy tools," reinforcing the potential impact of exposing these scripts. A malicious party can therefore gain an advantage by actively manipulating the rankings of popular domains. As we will show in the next section, such manipulation is actually feasible across all four lists, usually even on a sufficiently large scale without the need for significant resources.

3.5 Feasibility of large-scale manipulation

The data collection processes of popularity rankings rely on a limited view of the Internet, either by focusing on one specific metric or because they obtain information from a small population. This implies that targeted small amounts of traffic can be deemed significant on the scale of the entire Internet and yield good rankings. Moreover, the ranking providers generally do not filter out automated or fake traffic, or domains that do not represent real websites, further reducing the share of domains with real traffic in their lists.

Consequently, attacks that exploit these limitations are especially effective at allowing arbitrary modifications of the rankings at a large scale. We showed how adversaries may have incentives to skew the conclusions of security studies, and that security researchers and practitioners often use popularity rankings to drive the evaluation of these studies. Manipulating these rankings therefore becomes a prime vector for influencing security research, and as we will show, the small costs and low technical requirements associated with this manipulation make this approach even more attractive.

For each of the four studied popularity rankings, we describe techniques that manipulate the data collection process through the injection of forged data. To prove their feasibility, we execute those techniques that conform to our ethical framework and that have a reasonable cost, and show which ranks can be achieved. In Table 3.3, we summarize the techniques and the cost they incur on three aspects: money, effort and time required. Through this cost assessment, we identify how these manipulations could be applied at scale and affect a significant portion of these lists.

These techniques can be applied to both new domains and domains already present in the

Table 3.3: Summary of manipulation techniques and their estimated cost.

		Cost		
Provider	Technique	Monetary	Effort	Time
Alexa	Extension	none	medium	low
	Certify	medium	medium	high
Umbrella	Cloud providers	low	medium	low
Majestic	Backlinks	high	high	high
	Reflected URLs	none	high	medium
Quantcast	Quantified	low	medium	high

lists, e.g. when those domains bear the properties that could skew certain studies; a domain that has been ranked for a longer period of time may enjoy a higher trust or importance. In our work, we focus on techniques that directly influence the rankings' data at a modest cost. An alternative approach could be to buy expired or parked domains already in the list [347]. However, expired domains are usually bought up very quickly by "drop-catchers" [298], leaving a limited number of ranked domains open for registration [430]. Meanwhile, popular parked domains can command prices upwards of 1 000 USD [430]. This approach therefore incurs a prohibitive cost, especially at a large scale.

3.5.1 Alexa

Alexa ranks domains based on traffic data from two sources: their "Traffic Rank" browser extension that reports all page visits, and the "Certify" analytics service that uses a tracking script to count all visits on subscribing websites. We forge traffic data to both and observe the achieved ranks.

Extension

The "Alexa Traffic Rank" extension collects data on all pages that its users visit. The extension also shows users information on the rank and traffic of the visited site, which may serve as an incentive to install the extension.

We submitted page visits for both registered and nonexistent test domains previously unseen by Alexa. We generated profiles with all 1 152 possible configurations, i.e. the demographic details that are requested when installing the extension, and this within a short timeframe from the same IP address; Alexa did not impose any limits on the number of profiles that could be created. We submitted visits to one domain per profile; as visits to the same page by the same profile are only counted once [36], we generated exactly one visit per page to the homepage and randomly generated subpages. The number of page views for one test domain ranges from 1 to 30.

We installed the extension in a real Chrome browser instance and then generated page visits to our test domain, simulating a realistic usage pattern by spacing out page visits between 30 and 45 seconds, and interspersing them with as many visits to domains in Alexa's top 1000. Through inspection of the extension's source code and traffic, we found that upon page load, a GET request with the full URL of the visited page[9] is sent alongside the user's profile ID and browser properties to an endpoint on `data.alexa.com`. This means these requests can also be generated directly without the need to use an actual browser, greatly reducing the overhead in manipulating many domains on a large scale.

From May 10, 2018 onward, Alexa appears to block data reporting from countries in the European Union (EU) and European Economic Area (EEA), as the response changed from the visited site's rank data shown to the user to the string "Okay". This is likely due to the new General Data Protection Regulation coming into force. While we were able to circumvent this block through a VPN service, Alexa may be ignoring traffic in EU and EEA countries, introducing a further bias towards traffic from other countries.

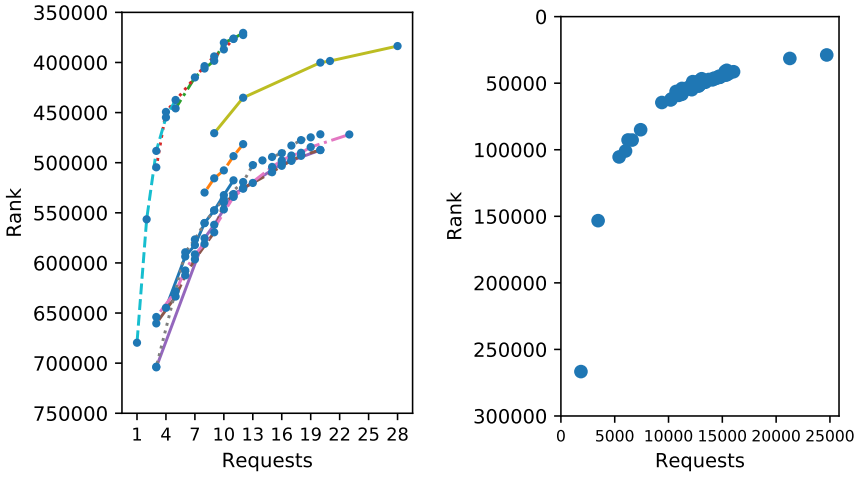
For 20% of our profiles/domains, we were successful in seeing our page views counted and obtaining rankings within the top million. Alexa indicates that it applies statistical processing to its data [528], and we suspect that some of our requests and generated profiles were pruned or not considered sufficient to be ranked, either because of the profile's properties (e.g. a common browser configuration or an overrepresented demographic) or because only a subset of traffic data is (randomly) selected. To increase the probability of getting domains ranked, an adversary can select only the successful profiles, or generate page views to the same site with different profiles in parallel, improving the efficiency of their manipulation.

Figure 3.6a lists our 224 successful rankings grouped per day, showing the relation between ranks and number of visits. We performed our experiments between July 25 and August 5, 2018. As during this period Alexa averaged traffic over one day, there was only a delay of one day between our requests and the domains being ranked; they disappeared again from the list the following day. This means that it is not necessary to forge requests over a longer period of time when the malicious campaign is short-lived.

What is most striking, is the very small number of page visits needed to obtain a ranking: as little as one request yielded a rank within the top million, and we achieved a rank as high as 370 461 with 12 requests (albeit in the week-end, when the same number of requests yields a better rank). This means that the cost to manipulate the rankings is minimal, allowing adversaries to arbitrarily alter the lists at large scale for an extended period of time. This ensures continued ranking and increases the likelihood of a list containing manipulated domains being used for research purposes, despite the large daily change.

The low number of required requests is further confirmed by large blocks of alphabetically ordered domains appearing in the ranking: these point towards the same number of visits being counted for these domains. We use these blocks as well as the processed visitor

[9] For pages loaded over HTTPS, the path is obfuscated.



(a) Extension.

(b) Certify.

Figure 3.6: Ranks obtained in the Alexa list. Ranks on the same day are connected.

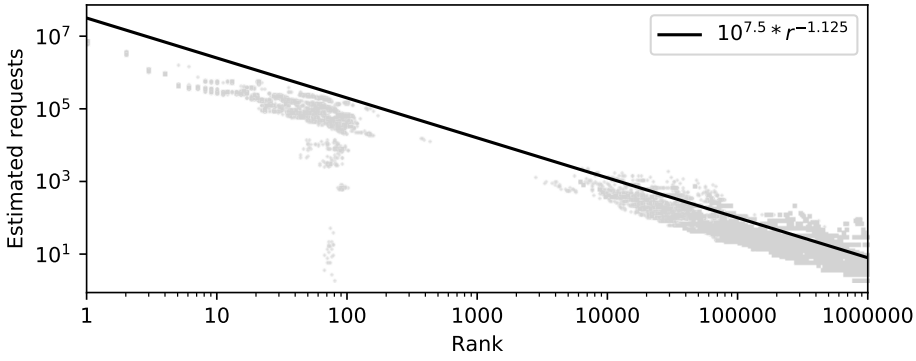


Figure 3.7: The estimated relation between requests and rank for Alexa. The gray areas show data as retrieved from the Alexa Web Information Service.

and view metrics retrieved from the Alexa Web Information Service [49] to estimate the required visit count for better ranks.

Figure 3.7 shows the number of requests needed to achieve a certain rank; we consider this an upper bound as Alexa ranks domains that see more unique visitors better than those with more page views, meaning that manipulation with multiple profiles would require less requests. This analysis shows that even for very good ranks, the amount of requests required and accompanying cost remains low, e.g. only requiring 1 000 page views for rank 10 000. This model of Alexa's page visits also corresponds with previous observations of Zipf's law in web traffic [21, 121].

Alexa's list is also susceptible to injection of nonexistent domains; we were able to enter one such domain. Furthermore, we confirmed in our server logs that none of our test domains were checked by Alexa as we forged page visit requests. The ability to use fake domains reduces the cost to manipulate the list at scale even further: an attacker is not required to actually purchase domain names and set up websites for them.

Even though Alexa's statistical postprocessing may prune some visits, the low number of required visits, the ability to quickly generate new profiles and the lack of filtering of fake domains allows an attacker to still easily achieve significant manipulation of Alexa's list.

Certify

Alexa's 'Certify' service offers site owners an analytics platform, using a tracking script installed on the website to directly measure traffic. The service requires a subscription to Alexa's services, which start at USD 19.99 per month for one website.

As Alexa verifies installation of its scripts before tracking visits, we installed them on a test website. From the JavaScript part of this code, we extracted its reporting algorithm and repeatedly forged GET requests that made us appear as a new user visiting the website, therefore avoiding the need to retain the response cookies for continued tracking. To diversify the set of IP addresses sending this forged traffic, we sent these requests over the Tor network, which has a pool of around 1 000 IP addresses [473]. We sent at most 16 000 requests per 24 hours, of which half were for the root page of our domain, and the other half for a randomly generated path.

Figure 3.6b lists the ranks of our test domain and the number of visits that were logged by Alexa across 52 days. For 48 days, we reached the top 100 000 (purported to more accurately reflect popularity), getting up to rank 28 798. Not all our requests were seen by Alexa, but we suspect this is rather due to our setup (e.g. by timeouts incurred while sending requests over Tor). Alexa's metrics report that our site received "100.0% real traffic" and that no traffic was excluded, so we suspect that Alexa was not able to detect the automated nature of our requests.

After subscription to the service, Alexa will only calculate (and offer to display) the 'Certified' rank of a website after 21 days. Since no visits to our site were being reported through Alexa's extension, no 'normal' rank was achieved in the meantime, and therefore there was a large delay between the start of the manipulation and the ranking of the domain.

The disadvantage of this technique is that the cost of manipulation at scale quickly becomes prohibitive, as for each site that needs to be inserted into the list, a separate subscription is required. Given Alexa's verification of the tracking script being installed, the domain needs to be registered and a real website needs to be set up, further reducing the scalability of the technique. However, we were able to achieve better ranks with a more consistent acceptance of our forged requests. Depending on the attacker's goal, it is

of course still possible to artificially increase the ranking of specific websites who already purchased and installed the Alexa Certify service.

We obtained a rank even though we did not simulate traffic to this test domain through the Alexa extension, which strongly suggests that Alexa does not verify whether 'Certified' domains show similar (credible) traffic in both data sources. Based on this observation, we found one top 100 'Certified' site where Alexa reports its extension recording barely any or even no traffic: while in this case it is a side-effect of its usage pattern (predominantly mobile), it implies that manipulation conducted solely through the tracking script is feasible.

3.5.2 Cisco Umbrella

Umbrella ranks websites on the number of unique client IPs issuing DNS requests for them. Obtaining a rank therefore involves getting access to a large variety of IP addresses and sending (at least) one DNS request from those IPs to the two open DNS resolvers provided by Umbrella.

Cloud providers

Cloud providers have obtained large pools of IP addresses for distribution across their server instances; e.g. Amazon Web Services (AWS) owns over 64 million IPv4 addresses [50]. These can be used to procure the unique IP addresses required for performing DNS requests, but due to their scarcity, providers restrict access to IPv4 addresses either in number or by introducing a cost.

In the case of AWS, there are two options for rapidly obtaining new IPv4 addresses. Continuously starting and stopping instances is an economical method, as even 10 000 different IPs can be obtained for less than USD 1 (using the cheapest instance type), but the overhead of relaunching instances reduces throughput: on the cheapest `t2.nano` instance, we were able to obtain a new IP on average every minute. Moreover, the number of concurrent running instances is limited, but by using instances in multiple regions or even multiple accounts, more instances are accessible. Keeping one instance and allocating and deallocating Elastic IP addresses (i.e. addresses permanently assigned to a user) yields higher throughput, at 10 seconds per IP. However, AWS and other providers such as Microsoft Azure discourage this practice by attaching a cost to this 'remap' operation: for AWS, a remap costs USD 0.10, so a set of 10 000 IPs incurs a prohibitive cost of USD 1 000.

Figure 3.8 shows the relation between the number of issued DNS requests and the obtained rank; all of our attempts were successful. We were able to obtain ranks as high as 200 000 with only a thousand unique IP addresses, albeit in the weekend, when OpenDNS processes around 30% less DNS traffic [367]. We only sustained DNS traffic for one day at a time, but it appears that Umbrella counts this traffic (and therefore ranks the domain)

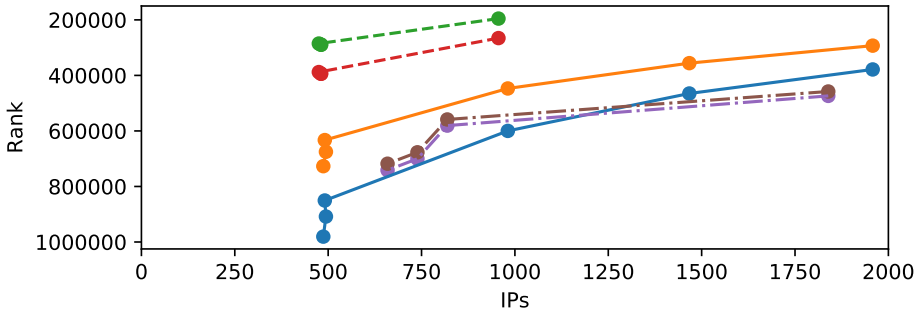


Figure 3.8: Ranks obtained in the Umbrella list. Ranks on the same day are connected; ranks over two days for one set of requests use the same line style.

for two days, reducing the number of requests needed per day to either obtain a good rank for one domain or rank many domains.

Given the relatively high cost per IP, inserting multiple domains actually is more economical as several DNS requests can be sent for each IP instantiation. As the name requested in the DNS query can be chosen freely, inserting fake domains is also possible; the high number of invalid entries already present shows that Umbrella does not apply any filtering. This further improves scalability of this technique, as no real websites need to be set up in order to manipulate the list.

The effort to generate many ranked entries is further reduced by the inclusion of subdomains, as all subdomains at lower depths are automatically ranked: we were able to rank 12 subdomains simultaneously with one set of requests. Furthermore, the number of requests is aggregated per subdomain, so a low number of requests to many subdomains can result in both many ranked subdomains and a good rank for the pay-level domain.

Combining the ability to insert fake domains with the low overhead of requests to additional domains, the inclusion of subdomains and the lack of any filtering or manipulation detection means that the scale at which an attacker can manipulate Umbrella's list can be very large.

Alternatives

Tor. The Tor service provides anonymous communication between a user and the service they use. Traffic is relayed across multiple nodes before being sent to the destination from an exit node, meaning that the destination observes traffic originating from that node's IP address. This set of exit nodes provide a pool of IP addresses, and by switching the routing over the Tor network, DNS requests can be altered to appear to originate from multiple IP addresses in this pool. However, as there are less than 1 000 exit nodes at any given point in time [473], it will be possible to inject domains in the list, but infeasible to obtain a high rank solely through this technique.

IP spoofing. IP packets contain the IP address of its sender, that can however be arbitrarily set in a technique known as IP spoofing. We could leverage this technique to set the source IP of our DNS packets to many different addresses, in order for our requests to appear for Umbrella to originate from many unique IPs. As IP spoofing is often used during denial-of-service attacks, many ISPs block outgoing packets with source IPs outside their network. Leveraging IP spoofing for sending DNS requests therefore requires finding a network that supports it. Karami et al. [257] found that certain VPS providers allow IP spoofing; as such these could be used for our experiment.

Due to the ethical concerns that are raised by leveraging IP spoofing (the responses of our DNS requests would arrive at the users of the forged source IPs, and the associated traffic may cause the VPS provider to be flagged as malicious), we did not further explore this technique. It is important to note however that an adversary only needs to find a single provider or network that does not prevent IP spoofing in order to send a very large number of DNS requests to Umbrella's resolvers and thus manipulate the list at a very large scale.

3.5.3 Majestic

Majestic's ranking is based on the number of subnets hosting a website that links to the ranked domain. Therefore, we cannot construct data reporting requests sent directly to Majestic, but must use techniques where website owners knowingly or unknowingly serve a page that contains a link to our domain and that is then crawled independently by Majestic.

Backlinks

Backlink providers offer a paid service where they place incoming links for a requested website ('backlinks') on various sites. The goal of this service is usually to achieve a higher position in search engine rankings, as part of search engine optimization (SEO) strategies; the deceptive nature of this technique makes that this is considered 'black-hat' SEO.

Backlinks are priced differently according to the reputation of the linking site. While we need a sufficiently diverse set of websites hosted on different subnets, Majestic does not take the quality of our backlinks into account when ranking domains. This means that we can reduce our cost by choosing the cheapest type of backlink. Moreover, we have the choice of removing backlinks after they have been found, as these are no longer billed but still count towards the subnets for a period of at most 120 days, reducing monetary cost.

We use the services of `BackLinks.com`, as they operate only on sites under their control, therefore avoiding impact of our experiment on unaware site owners. The choice for this particular backlink provider brings about certain constraints (such as the pool of available backlink sites, or a limit on daily backlink deletions), but these can be alleviated by using other and/or multiple backlink providers. We buy backlinks if they are located in

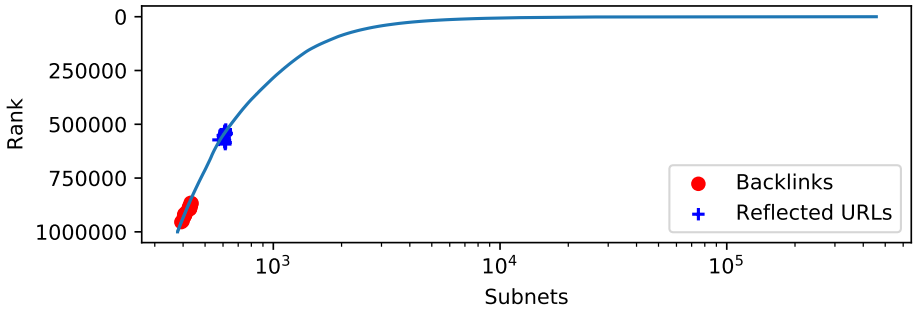


Figure 3.9: The relation between subnets and rank in the Majestic list for May 31, 2018, with our obtained ranks highlighted.

a subnet not covered by any already purchased site, but have to use OCR as the URLs on which links would be placed are only available as a warped image. We therefore curated the set of backlinks through manual verification to compensate for any errors, increasing our required effort.

The cheapest type of backlink costs USD 0.25 a month, but since there was not a sufficient amount of such pages to cover the necessary number of subnets, more expensive backlinks were also required. The backlinks were partially found organically by Majestic; in this case there is no additional cost. Through a subscription on Majestic's services, backlinks can also be submitted explicitly for crawling: the minimum cost is USD 49.99 for one month.

We bought backlinks for our test domain and curated them for two and a half months, in order to capture as many subnets as possible while managing the monetary cost. Our total cost was USD 500. We successfully inserted our domain, with Figure 3.9 showing the achieved rankings on top of the relation between the rank and the number of found subnets for all ranked sites as published by Majestic.

There exists a trade-off between the cost and the time required to enter the rank: if the monetary cost should be kept low, more time is needed as the set of eligible backlink pages is smaller and backlinks will need to be deleted. Alternatively, a higher number of possibly more expensive backlinks would allow to achieve the necessary number of subnets more quickly, but at a higher monetary cost. Conversely, because Majestic considers links for at least 120 days, the cost for long-term manipulation is relatively limited: even though we stopped buying backlinks and these subsequently disappeared, our ranking was still maintained for more than two months as previously found backlinks were still counted.

Reflected URLs

An alternative technique that we discovered, for which it is not required to purchase services from external parties, is to leverage websites that reflect a GET parameter into

a link. Note that for our purpose, reflected cross-site scripting (XSS) attacks could also be used; however, this technique is more intrusive as it will inject HTML elements, so we did not evaluate it out of ethical considerations. To discover web pages that reflect a URL passed as a parameter, we started crawling the 2.8 million domains from the four lists, finding additional pages by following links from the homepage of these domains. If GET parameters were found on the page, we replaced each one with a URL and tested whether this URL was then included in the `href` of an `a` tag on the page.

Through this crawl, we found that certain MediaWiki sites were particularly susceptible to reflecting URLs on each page, depending on the configuration of the site. We therefore tested this reflection on the wikis from a number of data sources: the root domains as well as the subdomains containing `wiki` of the four top lists, the set of wikis found by Pavlo and Shi in 2011 [378] and the wikis found by WikiTeam[10]. As the reflection is purely achieved through altering the GET parameters, we do not permanently alter the wiki.

Given the special construction of their URLs, the pages reflecting our domain will not be found organically by Majestic. The list of affected URLs can be submitted directly to Majestic, but this requires a subscription. The links can also be placed on one aggregating web page: by verifying ownership of the hosting domain with Majestic, a crawl of this page and subsequently of the links placed on it can be triggered for free; alternatively, using Majestic's site to request the freely available subset of backlinks data for this special web page also seems to trigger this crawl.

Through our crawls, we found 1 041 pages that reflected the URL of our test domain when passed in a GET parameter. Through submitting these reflecting URLs to Majestic's crawler, we successfully ranked our domain, with Figure 3.9 showing the achieved rankings over time. Through this technique, we also successfully had one backlink to a non-existing domain crawled and counted as a referring subnet. By scaling this up to the number of subnets required to be ranked, this implies that Majestic's list ranking is also susceptible to fake entries; as there are unavailable sites in the list, Majestic likely does not actively check whether entries in the list are real.

This technique allows to construct backlinks at no monetary cost, but requires a high effort to find appropriate pages. We found only small subsets of wikis and domains in general to reflect our URL, so the number of pages and subnets that can be discovered using this technique may not be sufficient to achieve very high rankings. Given a deeper crawl of pages, more sites that reflect URLs passed through a GET parameters may be found, more subnets can be covered and a higher ranking can be achieved. Moreover, an attacker can resort to more 'aggressive' techniques where URLs are permanently stored on pages or XSS vulnerabilities are exploited.

Once found however, a reflecting URL will be counted indefinitely: a site would effectively have to be reconfigured or taken offline in order for the backlink to disappear. This means maintaining a rank comes at no additional cost. Furthermore, every website that is susceptible to URL reflection can be leveraged to promote any number of attacker-chosen

[10] `https://github.com/WikiTeam/wikiteam`

(fake) domains, at the cost of submitting more (crafted) URLs to Majestic. This means that manipulation of Majestic's list is also possible on a large scale.

Alternatives

Hosting own sites. Using domains seen in passive DNS measurements, Tajalizadehkhoob et al. [469] identified 45 434 hosting providers in 2016, and determined their median address space to contain 1 517 IP addresses. Based on these figures, we can assume that the number of subnets available through hosting providers is well above the threshold to be ranked by Majestic. An attacker could therefore set up websites on a sufficient number of these providers, all with a link back to the domain to be ranked. By making all the websites link to each other, a larger set of domains could easily be ranked. This technique incurs a high cost however: in effort, as setting up accounts with these providers is very likely to require a lot of manual effort, as well as in monetary cost, as for each hosting provider a subscription needs to be bought.

Pingbacks. Content management systems such as WordPress provide a pingback mechanism for automatically reporting URLs that link to one of the pages hosted on that system. Many sites will then insert a link back to the reported URL on that page. By finding a set of domains supporting pingbacks (similar to finding wikis) and reporting a URL on the domain we want to see ranked, we could again have links to our domain on a large set of domains and therefore subnets. However, this permanently changes pages on other websites, and although enabling the pingback feature implies some consent, we opted to not explore this technique for ethical reasons.

3.5.4 Quantcast

Quantified

Quantcast mainly obtains traffic data through its tracking script that webmasters install on their website. We extracted the reporting algorithm from the tracking script, and automatically sent requests to Quantcast from a set of 479 VPN servers located in the United States, as Quantcast's ranking only takes US traffic into account. We sent requests for 400 generated users per day, presenting ourselves as a new user on the first request and subsequently reusing the generated token and received cookie in four more requests. As opposed to Alexa's tracking script, reporting page views for only new users did not result in any visits being counted.

Our forged requests were acknowledged by Quantcast and its analytics dashboard reports that on May 30, 2018, "the destination reaches over 6,697 people, of which 6,696 (100%) are in the U.S." The latter metric is used to determine the rank. However, our test domain has not appeared in the ranking. This is likely due to the short age of our domain; although

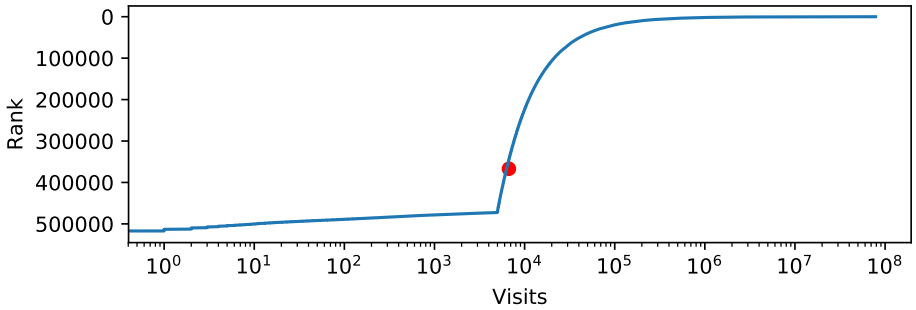


Figure 3.10: The relation between measured visits and rank in the Quantcast list for May 31, 2018, with the theoretical rank for our visit count highlighted.

we have sent requests for more than a month, Quantcast's slow update frequency means its ranking algorithm may not take our domain into account yet.

As Quantcast publishes the number of visits counted for each ranked domain, the relation between the desired rank and required effort is known as shown in Figure 3.10. Up to around 5 000 visits, the achieved rank remains relatively low; this tail contains primarily quantified sites that are ranked even with almost no visits. Above 5 000 visits, Quantcast's list includes many more domains for which a rank is estimated; especially at worse ranks, large blocks of estimated domains are interspersed with quantified domains, so increasing the number of visits to jump across such a block gives a large improvement in rank. If a rank were to be assigned to our domain, we can determine that we would theoretically be given a rank around 367 000. Achieving higher ranks only requires submitting more forged requests, so the increased cost in time and effort is minimal.

Quantcast will only start processing traffic data once it has verified (through crawling) that its tracking pixel is present on the domain. It is therefore required to register the domain and set up a real website to manipulate the rankings, so scaling to multiple domains incurs a higher cost; Quantcast's analytics platform itself is free however, limiting the additional cost. As Quantcast performs the check only once, the domain and the website also do not need to be sustained. Merely registering for tracking may even suffice to be ranked: over 2 000 domains are ranked but reported to have 0 visits, with over half registered by DirectEmployers as discussed in Section 3.3.3.

Alternatives

Quantcast states that it also uses traffic data from 'ISPs and toolbar providers' [442]. ISPs sell traffic data to third parties [100], and Quantcast may be buying these services to generate the number of page visits and therefore the rank for non-quantified websites. However, we cannot determine which ISPs may be used. As for extensions, we were unable to discover any extensions reporting to a URL that was obviously related to Quantcast.

Ethical considerations Because our experiments may have a large impact on the reputation of the rankings as well as potentially affect third parties, we conduct an ethical review of our experimental methods. Such reviews have been advocated for by the academic community [376] and ensure that the potential damage inflicted is minimized. We base this review on the ethical principles outlined in the Menlo Report [146], which serves as a guideline within the field of ICT research; we apply the principle of beneficence in particular: identifying potential benefits and harms, weighing them against each other and minimizing the risk of inflicting harm.

Because of their commercial nature, the providers of popularity rankings have an economic interest in these being accurate. We show that these lists can be manipulated, negatively affecting their perceived reputability. Our findings are however of value to the providers: by evaluating the various techniques and reporting our findings, the providers become aware of the potential threats, may take actions to thwart attackers and can improve the correctness of their rankings.

We have disclosed our findings and proposals for potential remedies to the four providers, alongside a list of manipulated domains for them to remove from their datasets and past and present rankings. Alexa and Majestic provided statements regarding the value of their rankings and the (in)feasibility of manipulation, but commercial considerations prevent them from elaborating on their methods. Cisco Umbrella closed our issue without any statement, and we received no response from Quantcast. None of our test domains were (retroactively) removed from any rankings after our notification.

We minimize the impact of our experiments on third parties by only significantly manipulating the ranking of our own, purposefully registered domains and refraining from intrusive or questionable techniques. Our sites also contained an explanation of our experiment and contact details for affected parties. Our low number of test domains means that only few domains will see negligible shifts in ranking due to our experiments; e.g. the volatility of Alexa's list has a significantly larger impact. Moreover, we minimized the duration of our experiments and of our domains being ranked. The impact on other research using these lists is also minimal; we showed that in general many more ranked domains are unavailable or unrepresentative. Our sites only hosted benign content, so whitelists using rankings are unaffected.

3.6 An improved top websites ranking

As we showed, the different methods used to generate popularity rankings cause undesirable effects on their properties that can potentially sway the results and conclusions of studies. In addition, we showed that researchers are prone to ignore or be unaware of these effects. We also proved that these rankings show several pitfalls that leave them vulnerable to large-scale manipulation, further reducing their reliability and suitability to research. Nevertheless, popularity rankings remain essential for large-

scale empirical evaluations, so we propose improvements to existing rankings as well as a new ranking that has characteristics geared towards research.

3.6.1 Defending existing rankings against manipulation

Even though the methods for data collection and processing of the existing lists are usually unknown, our experiments suggest that their providers employ little defense against large-scale manipulation. We outline techniques that the providers could use to make these lists more resilient to attacks.

Detecting and deterring singular instances of fraud ensures that all data used in ranking domains is deemed valid. Alexa and Quantcast rely on the reporting of page visits; within the realm of online advertising, techniques have been designed to subvert click inflation [13, 90, 328]. As we saw that not all attempts at manipulating Alexa's ranking were successful, this may imply that Alexa already employs some of these tactics.

To deter large-scale manipulation, ranking providers could employ tactics that increase the effort and resources required to affect many domains to prohibitive levels. This therefore avoids significant influence on research results, even if these tactics may not be sufficient to stop small-scale manipulation.

For a traffic reporting extension, the profile setup could be tied to an account at an online service; while a normal user can easily create one account, creating many accounts in an automated way can be countered by techniques that try to detect fake accounts [105]. In the case of Alexa, given its ownership by Amazon, a natural choice would be to require an Amazon account; in fact, a field for such an account ID is available when registering the extension, but is not required. This technique is not useful for tracking scripts, since no user interaction can be requested, and fraud detection as discussed earlier may be required. For providers that use both, the two metrics can be compared to detect anomalies where only one source reports significant traffic numbers, as we suspect such manipulation is already happening for Alexa Certify.

Data could be filtered on the IP address from which it originates. Ignoring requests from ranges belonging to cloud providers or conversely requiring requests to come from ranges known to belong to Internet service providers (e.g. through its autonomous system) does not block a single user from reporting their traffic. However, using many IP addresses concurrently is prevented as these cannot be easily obtained within the permitted ranges. This technique is particularly useful for Umbrella's list; for the other lists, using many IP addresses is not strictly necessary for large-scale manipulation.

The relative difficulty of maliciously inserting links into pages on many IP subnets already reduces the vulnerability of link-based rankings to large-scale manipulation. Specific attacks where the page reflects a URL passed as a parameter could be detected, although this can be made more difficult by obfuscation and attacks that alter a page more permanently. The link-based rankings could be refined with reputation scores, e.g.

the age of a linked page or Majestic's "Flow Metrics" [315], to devalue domains that are likely to be part of a manipulation campaign.

Finally, requiring ranked domains to be available and to host real content increases the cost of large-scale manipulation, as domain names need to be bought and servers and web pages need to be set up. For Umbrella, not ranking domains where name resolution fails can significantly reduce unavailable (and therefore possibly fake) domains in the list. The other providers can perform similar availability checks in the DNS or by crawling the domain.

3.6.2 Creating rankings suitable for research

As we cannot ensure that providers will (want to) implement changes that discourage (large-scale) manipulation, we look at combining all currently available ranking data with the goal of improving the properties of popularity rankings for research, canceling out the respective deficiencies of the existing rankings. To this extent, we introduce Tranco, a service that researchers can use to obtain lists with such more desirable and appropriate properties. We provide standard lists that can be readily used in research, but also allow these lists to be highly configurable, as depending on the use case, different traffic sources or varying degrees of stability may be beneficial.

Moreover, we provide a permanent record to these new lists, their configuration and their construction methods. This makes historical lists more easily accessible to reduce the effort in replicating studies based upon them, and ensures that researchers can be aware of the influences on the resulting list by its component lists and configuration.

Our service is available at **`https://tranco-list.eu`**. The source code is also openly published at `https://github.com/DistriNet/tranco-list` to provide full transparency of how our lists are processed.

Combination options and filters

We support creating new lists where the ranks are averaged across a chosen period of time and set of providers, and introduce additional filters, with the goal of enhancing the research-oriented properties of our new lists.

In order to improve the rank of the domains that the lists agree upon, we allow to average ranks over the lists of some or all providers. We provide two combination methods: the Borda count where, for a list of length N, items are scored with $N, N-1, ..., 1, 0$ points; and the Dowdall rule where items are scored with $1, 1/2, ..., 1/(N-1), 1/N$ points [185]. The latter reflects the Zipf's law distribution that website traffic has been modeled on [21, 121]. Our standard list applies the Dowdall rule to all four lists. We also allow to filter out domains that appear only on one or a few lists, to avoid domains that are only marked as popular by one provider: these may point to isolated manipulation.

To improve the stability of our combined lists, we allow to average ranks over the lists of several days; our standard list uses the lists of the past 30 days. Again, we allow to filter out domains that appear only for one or a few days, to avoid briefly popular (or manipulated) domains. Conversely, if capturing these short-term effects is desired, lists based on one day's data are available. When combining lists, we also provide the option to only consider a certain subset of the input lists, to select domains that are more likely to actually be popular.

Differences in list composition complicate the combination of the lists. Umbrella's list includes subdomains; we include an option to use a recalculated ranking that only includes pay-level domains. Quantcast's list contains less than one million domains; we proportionally rescale the scores used in the two combination methods to the same range as the other lists.

We add filters to create a list that represents a certain desired subset of popular domains. A researcher can either only keep domains with certain TLDs to select sites more likely to be associated with particular countries or sectors, or exclude (overly represented) TLDs. To avoid the dominance of particular organizations in the list, a filter can be applied where only one domain is ranked for each set of pay-level domains that differ only in TLD. Finally, only certain subdomains can be retained, e.g. to heuristically obtain a list of authentication services by selecting `login.*` subdomains.

To allow researchers to work with a set of domains that is actually reachable and representative of real websites, we provide options to filter the domains on their responsiveness, status code and content length. We base these filters on a regular crawl of the union of all domains on the four existing lists. This ensures that the sample of domains used in a study yields results that accurately reflect practices on the web.

To further refine on real and popular websites, we include a filter on the set of around 3 million distinct domains in Google's Chrome User Experience Report, said to be 'popular destinations on the web' [116]. Its userbase can be expected to be (much) larger than e.g. Alexa's panel; however, Google themselves indicate that it may not fully represent the broader Chrome userbase [116]. Moreover, the list is only updated monthly and does not rank the domains, so it cannot be used as a replacement for the existing rankings.

To reduce the potential effects of malicious domains on research results (e.g. in classifier accuracy), we allow to remove domains on the Google Safe Browsing list [418] from our generated lists.

Evaluation

We evaluate the standard options chosen for our combined lists on their improvements to similarity and stability; the representativeness, responsiveness and benignness of the included domains can be improved by applying the appropriate filters. We generate our combined lists from March 1, 2018 to November 14, 2018, to avoid distortions due to

Alexa's and Quantcast's method changes, and truncate them to one million domains, as this is the standard for current lists.

Similarity To determine the weight of the four existing lists, we calculate the rank-biased overlap with our combined lists. Across different weightings, the RBO with Alexa's and Majestic's lists is highest at 46.5–53.5% and 46.5–52% respectively, while the RBO with Quantcast's and Umbrella's lists is 31.5–40% and 33.5–40.5% respectively. These results are affected by the differences in list composition: subdomains for Umbrella and the shorter list for Quantcast mean that these two lists have less entries potentially in common with Alexa and Majestic, reducing their weight. Overall, there is no list with a disproportionate influence on the combined list.

Stability Averaging the rankings over 30 days is beneficial for stability: for the list combining all four providers, on average less than 0.6% changes daily, even for smaller subsets. For the volatile Alexa and Umbrella lists, the improvement is even more profound: the daily change is reduced to 1.8% and 0.65% respectively. This means that the data from these providers can be used even in longitudinal settings, as the set of domains does not change significantly.

Reproducibility

Studies rarely mention the date on which a ranking was retrieved, when the websites on that list were visited and whether they were reachable. Moreover, it is hard to obtain the list of a previous date: only Cisco Umbrella maintains a public archive of historical lists [119]. These two aspects negatively affect the reproducibility of studies, as the exact composition of a list cannot be retrieved afterwards.

In order to enhance the reproducibility of studies that use one of our lists, we include several features that are designed to create a permanent record that can easily be referenced. Once a list has been created, a permanent short link and a preformatted citation template are generated for inclusion in a paper. Alongside the ability to download the exact set of domains that the list comprises, the page available through this link provides a detailed overview of the configuration used to create that particular list and of the methods of the existing rankings, such that the potential influences of the selected method can be assessed. This increases the probability that researchers use the rankings in a more well founded manner.

Manipulation

Given that our combined lists still rely on the data from the four existing lists, they remain susceptible to manipulation. As domains that appear on all lists simultaneously

are favored, successful insertion in all lists at once will yield an artificially inflated rank in our combined list.

However, the additional combinations and filters that we propose increase the effort required to have manipulated domains appear in our combined lists. Averaging ranks over a longer period of time means that manipulation of the lists needs to be maintained for a longer time; it also takes longer for the manipulated domains to obtain a (significant) aggregated rank. Moreover, intelligently applying filters can further reduce the impact of manipulation: e.g. removing unavailable domains thwarts the ability to use fake domains.

As each ranking provider has their own traffic data source, the effects of manipulating one list are isolated. As none of the lists have a particularly high influence in the combined list, all four lists need to manipulated to the same extent to achieve a comparable ranking in the combined list, quadrupling the required effort. For the combined list generated for October 31, 2018, achieving a rank within the top million would require boosting a domain in one list to at least rank 11 091 for one day or rank 332 778 for 30 days; for a rank within the top 100 000, ranks 982 and 29 479 would be necessary respectively. This shows that massive or prolonged manipulation is required to appear in our combined list.

3.7 Related work

The work that is most recent and most closely related to ours is that of Scheitle et al. [427], who compared Alexa's, Majestic's and Umbrella's lists on their structure and stability over time, discussed their usage in (Internet measurement) research through a survey of recent studies, calculated the potential impact on their results, and drafted guidelines for using the rankings. We focus on the implications of these lists for security research, expanding the analysis to include representativeness, responsiveness and benignness. Moreover, we are the first to empirically demonstrate the possibility of malicious large-scale manipulation, and propose a concrete solution to these shortcomings by providing researchers with improved and publicly available rankings.

In 2006, Lo and Sedhain [307] studied the reliability of website rankings in terms of agreement, from the standpoint of advertisers and consumers looking for the most relevant sites. They discussed three ranking methods (traffic data, incoming links and opinion polls) and analyzed the top 100 websites for six providers, all of which are still online but, except for Alexa, have since stopped updating their rankings.

Meusel et al. [329] published one-time rankings of websites[11], based on four centrality indices calculated on the Common Crawl web graph [124]. Depending on the index, these ranks vary widely even for very popular sites. Moreover, such centrality indices can be affected by manipulation [215, 363].

In his analysis of DNS traffic from a Tor exit node, Sonntag [454] finds that popularity according to Alexa does not imply regular traffic over Tor, listing several domains with a

[11] http://wwwranking.webdatacommons.org/

good Alexa rank but that are barely seen in the DNS traffic. These conclusions confirm that different sources show a different view of popularity, and that the Alexa list may not be the most appropriate for all types of research (e.g. into Tor).

3.8 Conclusion

We find that 133 studies in recent main security conferences base their experiments on domains from commercial rankings of the 'top' websites. However, the data sources and methods used to compile these rankings vary widely and their details are unknown, and we find that hidden properties and biases can skew research results. In particular, through an extensive evaluation of these rankings, we detect a recent unannounced change in the way Alexa composes its list: their data is only averaged over a single day, causing half of the list to change every day. Most probably, this unknowingly affected research results, and may continue to do so. However, other rankings exhibit similar problems: e.g. only 49% of domains in Umbrella's list respond with HTTP status code 200, and Majestic's list, which Quad9 uses as a whitelist, has more than 2 000 domains marked as malicious by Google Safe Browsing.

The reputational or commercial incentives in biasing the results of security studies, as well as the large trust placed in the validity of these rankings by researchers, as evidenced by only two studies putting their methods into question, makes these rankings an interesting target for adversarial manipulation. We develop techniques that exploit the pitfalls in every list by forging the data upon which domain rankings are based. Moreover, many of these methods bear an exceptionally low cost, both technically and in resources: we only needed to craft a single HTTP request to appear in Alexa's top million sites. This provides an avenue for manipulation at a very large scale, both in the rank that can be achieved and in the number of domains artificially inserted into the list. Adversaries can therefore sustain massive manipulation campaigns over time to have a significant impact on the rankings, and, as a consequence, on research and the society at large.

Ranking providers carry out few checks on their traffic data, as is apparent from our ability to insert nonexistent domains, further simplifying manipulation at scale. We outline several mitigation strategies, but cannot be assured that these will be implemented. Therefore, we introduce Tranco, a new ranking based on combining the four existing lists, alongside the ability to filter out undesirable (e.g. unavailable or malicious) domains. These combined lists show much better stability over time, only changing by at most 0.6% per day, and are much more resilient against manipulation, where even manipulating one list to reach the top 1 000 only yields a rank of 100 000 in our combined list. We offer an online service at `https://tranco-list.eu` to access these rankings in a reproducible manner, so that researchers can continue their evaluation with a more reliable and suitable set of domains. This helps them in assuring the validity, verifiability and reproducibility of their studies, making their conclusions about security on the Internet more accurate and well founded.

4

Evaluating the Long-term Effects of Parameters on the Characteristics of the Tranco Top Sites Ranking

This chapter is based on the homonymous paper published in the proceedings of the 12th USENIX Workshop on Cyber Security Experimentation and Test (CSET 2019) [289]. This work was co-authored with Tom Van Goethem and Wouter Joosen.

Although researchers often use top websites rankings for web measurements, recent studies have shown that due to the inherent properties and susceptibility to manipulation of these rankings, they potentially have a large and unknown influence on research results and conclusions. As a response, we provide Tranco [291], a research-oriented approach for aggregating these rankings transparently and reproducibly.

We analyze the long-term properties of the Tranco ranking and determine whether it contains a balanced set of domains. We compute how well Tranco captures websites that are responsive, regularly visited and benign. Through one year of rankings, we also examine how the default parameters of Tranco create a stable, robust and comprehensive ranking.

Through our evaluation, we provide an understanding of the characteristics of Tranco that are important for research and of the impact of parameters on the ranking composition. This informs researchers who want to use Tranco in a sound and reproducible manner.

4.1 Introduction

When measuring the prevalence of (security) practices and issues or evaluating novel tools and approaches across the web, researchers often rely on rankings of the most popular websites to obtain a representative sample of domains for their study, most often the Alexa top 1 million ranking [33]. Even though these rankings are widely used in research, the methods for composing these rankings are opaque, not well-known and rarely questioned by the research community. In fact, several recent unannounced changes in ranking availability [39] and methods [291] highlight how dependent web-related research is on these rankings, and how these may unknowingly have a large impact on research results and conclusions.

Only recently have researchers started to examine the rankings and their potential influences on Internet measurement and web security research, finding flaws in their inherent properties and susceptibility to malicious manipulation, in particular the most widely used Alexa ranking [291, 416, 427]. However, access to reliable and representative lists of domains remains important, as it enables researchers to study the web in a sound manner.

In prior work, we therefore proposed Tranco [291], a ranking that is oriented towards research by providing improved characteristics, a transparent method and reproducible rankings. This ranking aggregates four existing rankings (Alexa, Majestic, Quantcast and Umbrella) over a customizable period of time, and allows to apply filters that tailor the list to a researcher's needs. Both a daily updated ranking that uses default parameters and a service to generate custom rankings are available to researchers at `https://tranco-list.eu/`.

Top sites rankings should contain a sufficient number of websites that are available for crawling without errors, and that can actually be considered as popular. Ideally, such rankings should also capture changes in popularity over time while still being sufficiently stable as to enable longitudinal studies. While previous work assessed these desirable properties for the four existing rankings, such an extensive analysis of the Tranco ranking has not yet been done.

We now evaluate Tranco similarly to the other rankings, in order to determine whether it is a valid research-oriented alternative to the currently used rankings. Moreover, using the openly available source code, we analyze the parameters available in Tranco, such as the aggregation period or scoring method, and assess their impact on the ranking's composition. Finally, we assess whether the default parameters selected for the daily updated Tranco ranking provide a suitable set of domains that can be used in a broad set of studies.

Our contributions are the following: 1) we generate the Tranco ranking over one year to evaluate its long-term properties, 2) we quantify the unresponsive and malicious sites to inform researchers' assumptions of domain characteristics [291], 3) we compare Tranco with existing popularity rankings, finding a larger overlap with more stable lists and a

good overlap with observed web traffic, and 4) we find that the default parameters provide a stable and consistent ranking.

4.2 Methods of the Tranco ranking

The Tranco ranking aggregates four existing top sites rankings, that all use different vantage points and data collection periods to compute the scores for their ranking [291]:

- Alexa[1] ranks 1 million mainly 'pay-level domains'[2] based on web traffic collected either from users of its 'Alexa Traffic Rank' browser extension or from website visitors through an analytics script. Ranks are based on 1 day of data.
- Majestic[3] ranks 1 million mainly 'pay-level domains' based on incoming links collected through a web crawl. Ranks are based on 120 days of data.
- Quantcast[4] ranks around 500,000 mainly 'pay-level domains' based on web traffic collected either from website visitors through an analytics script or from ISP data. Ranks are based on 30 days of data.
- Umbrella[5] ranks 1 million 'pay-level domains' and subdomains based on DNS traffic collected through its OpenDNS resolvers. Ranks are based on 2 days of data.

The aggregate score of each domain in the Tranco ranking is calculated as the sum of scores across all rankings within the aggregation period. These individual scores are derived from the rank value in a component ranking: either through the Borda method, where the score is the total number of items minus the rank, or the Dowdall method, where it is the inverse of the rank [185].

Using this scoring method, any subset of the four providers can be aggregated over any aggregation period. However, in order to offer an readily available, easy to use and consistent ranking to researchers, a standard ranking with default settings is generated daily and published online[6]. In this default ranking, all four rankings are aggregated over a period of 30 days, with domains being scored with the Dowdall method.

In order to support studies that require domains with specific properties, certain filters can be applied to obtain a set of appropriate domains. Domains can be filtered on their components: only keeping 'pay-level domains', certain TLDs, certain subdomains (e.g. only `login`) and/or one domain with a certain second-level label (e.g. only one `google.*`

[1] `https://www.alexa.com/topsites`

[2] A pay-level domain is a domain name that a consumer or business can directly register, and consists of a subdomain of a public suffix or effective top-level domain [349] (e.g. `.com` but also `.co.uk`).

[3] `https://majestic.com/reports/majestic-million`

[4] `https://www.quantcast.com/top-sites/`

[5] `https://umbrella-static.s3-us-west-1.amazonaws.com/index.html`

[6] `https://tranco-list.eu/`

domain). Moreover, a researcher can require domains to appear in the rankings of a certain number of providers or for a certain number of days. Finally, domains can be checked against other lists, such as the Chrome User Experience Report [116] or Google Safe Browsing [418], to retain only regularly visited websites (Section 4.3.2) or domains that have not been flagged as malicious (Section 4.3.5) respectively. In the default Tranco ranking, only pay-level domains are retained.

All domains in the union of the aggregated rankings receive a score. As the component rankings contain some domains unique to them, this union is usually larger than one million domains. The default ranking is truncated to one million domains, in line with existing rankings. The larger union also means that after filtering, there usually still remain at least one million domains that satisfy the applied filters.

4.3 Analysis of Tranco's properties

To evaluate the properties of the Tranco ranking, we generate rankings following the method described in our prior work [291], using the publicly available source code[7] and an archive of Tranco's component rankings. Even though (custom) rankings can be generated online, we generate them separately to reduce the burden on this service and to reduce processing time by parallelizing and optimizing the generation (e.g. generating 14-day lists from pairs of 7-day lists).

Our analyses are based on those that we previously conducted on the four rankings that constitute Tranco [291]. When we describe a ranking for a date D with an aggregation period of N days, this ranking aggregates the component rankings from $N-1$ days before D until and including D. Unless otherwise mentioned, we use the default parameters of the Tranco ranking: all four component rankings, an aggregation period of 30 days, the Dowdall method for scoring domains, only retaining pay-level domains and truncating the ranking at one million domains. We calculate aggregate scores over 1 day starting from April 1, 2018; as we construct rankings aggregated over longer periods from those calculated for shorter periods, we generate the default 30-day rankings starting from April 30, 2018 until April 30, 2019.

4.3.1 Similarity with component rankings

The four component rankings of Tranco have different properties (e.g. stability over time) due to the differences in data sourcing and processing (e.g. the data collection period). Even though Tranco considers its component rankings equally when calculating global domain scores, these differences can cause the rankings to have a different similarity with and influence on Tranco. Moreover, the resulting differences in ranking composition may also affect other properties, such as the responsiveness of domains in Tranco. In this

[7] https://github.com/DistriNet/tranco-list

section, we analyze the similarity of the component rankings with Tranco; throughout the rest of this paper, we analyze the contribution of each component ranking to the other properties of Tranco.

Table 4.1 shows the rank-biased overlap [513] (a similarity measure where better ranks receive a higher weight, configured through a parameter p) between Tranco and its component rankings, averaged over April 2019. Alexa and Majestic have a similar overlap with Tranco, at 55.0% and 55.8% respectively when the top 100,000 is weighted at 85.2%. Quantcast is third at 43.6%: its shorter list means that it can contribute fewer domains to Tranco. Finally, Umbrella has a low overlap at 14.4%: this can be attributed to the subdomains in Umbrella that are not retained in the default Tranco ranking. In general, overlap improves when a smaller subset of the rankings is more heavily weighted (lower p), showing that rankings tend to agree more on the head of the ranking but less on the long tail.

Figure 4.1 shows in detail which subsets of the four component rankings contribute to the Tranco ranking on April 30, 2019[8]. The Majestic ranking has the highest intersection of 627,341 domains, with its top 500,000 being almost evenly represented throughout the full Tranco ranking. Quantcast shares 338,588 domains (70.32% of its domains), with a high influence on the very top of the Tranco ranking. Alexa shares 421,916 domains, which are mostly well ranked; even though it shares about one third of domains less than Majestic, its contribution to the better ranked domains in Tranco translates into a similar rank-biased overlap. Finally, Umbrella shares only 165,244 domains (all pay-level) across all its one million domains, almost exclusively with the Tranco top 200,000.

We see that more stable rankings, i.e. Majestic and Quantcast, have a higher overlap with Tranco. Given that the four component rankings do not tend to agree on which domains are the most popular [291], we estimate that this higher overlap is not due to one ranking containing more domains that are also included in any of the other rankings. Instead, this is due to domains in more stable lists being ranked repeatedly (i.e. over a longer period) and therefore receiving a higher aggregate score. A more evenly distributed contribution of each ranking can therefore be achieved by reducing the aggregation period. Overall, while some rankings may have a higher contribution due to their stability, each component ranking contributes to some extent to the Tranco ranking.

4.3.2 Comparison with web traffic

The Chrome User Experience Report is a data set released by Google that contains website performance metrics for over 4.3 million distinct domains (May 2019) [116]. These domains are said to be 'popular destinations on the web', having been observed sufficiently regularly in Chrome user traffic. While the report is not designed to be a ranking, we can still use it to assess whether Tranco contains domains that are regularly visited in the currently most popular browser [359]. This would mean that Tranco

[8] Available at `https://tranco-list.eu/list/7PJX/1000000`.

Table 4.1: Average rank-biased overlap (RBO) in April 2019 between Tranco and its component rankings for five values of the parameter p, i.e. weightings of subsets.

	Subset with weight		Rank-biased overlap of Tranco and			
RBOp	85.2%	99.9996%	Alexa	Majestic	Quantcast	Umbrella
0.9	10	100	67.18%	70.32%	54.41%	25.92%
0.99	100	1K	60.65%	54.73%	40.61%	23.21%
0.999	1K	10K	56.40%	54.01%	36.17%	17.36%
0.9999	10K	100K	56.02%	56.66%	40.42%	14.46%
0.99999	100K	1M	54.95%	55.76%	43.55%	14.42%

Table 4.2: Average overlap in April 2019 between the Chrome User Experience Report and top sites rankings.

Subset	Alexa	Majestic	Quantcast	Umbrella	Tranco
10	100%	90%	100%	33%	100%
100	94.7%	76.4%	100.0%	21.1%	91.6%
1K	93.33%	85.51%	98.54%	14.36%	90.73%
10K	92.54%	85.60%	96.19%	11.04%	90.60%
100K	87.13%	57.05%	94.71%	11.63%	82.47%
1M	68.12%	40.45%	76.90%	13.96%	62.24%

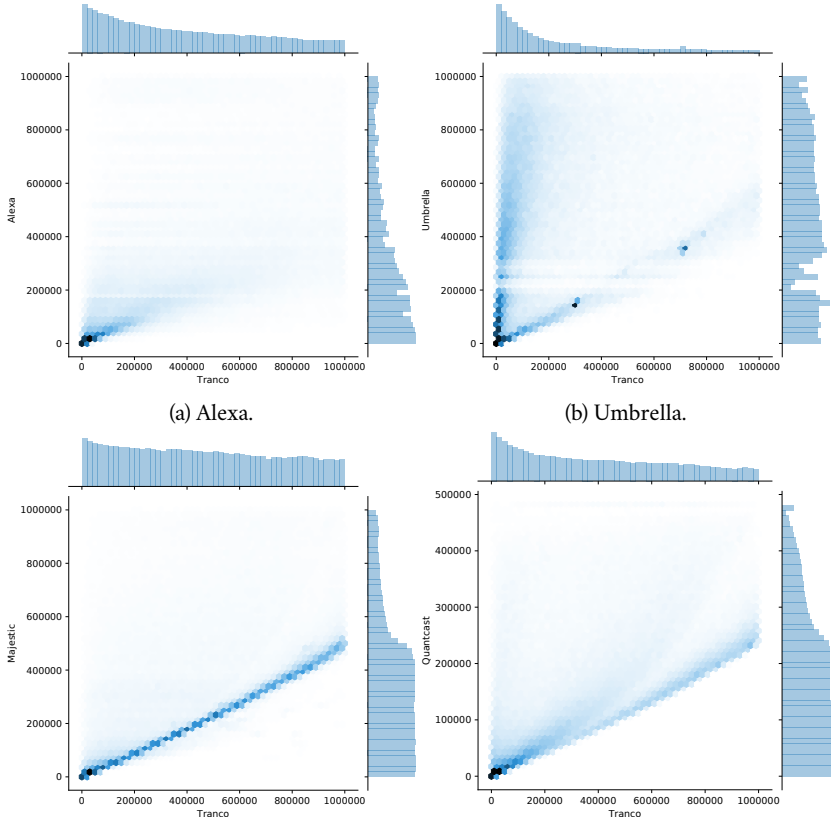
contains a representative sample of actual websites, and can therefore reliably be used to comprehensively study ecosystems on the web.

Table 4.2 shows that in April 2019, on average 62.24% of domains in the Tranco ranking were included in the Chrome User Experience Report of April 2019. This places Tranco after Quantcast and Alexa but before Majestic and Umbrella. Table 4.2 shows that overlap improves for smaller subsets of the ranking, with over 90% of the top 10,000 domains appearing in the Chrome User Experience Report.

We find that almost two thirds of domains in the Tranco ranking are sufficiently visited by Chrome users to be included in the Chrome User Experience Report. This indicates that the majority of Tranco's domains reflect real web traffic. Moreover, the domains in the Tranco ranking can be filtered upfront on whether they appear in the Chrome User Experience Report, guaranteeing that the resulting list consists only of actually visited websites.

4.3.3 Stability over time

The stability of a ranking determines how much the retrieval date impacts the obtained set of domains, which is particularly important for longitudinal studies. Stable rankings



(a) Alexa.

(b) Umbrella.

(c) Majestic.

(d) Quantcast.

Figure 4.1: Contribution of component rankings to the Tranco ranking of April 30, 2019. The right-hand histogram shows which domains from the component ranking appear in Tranco, while the top histogram shows at which ranks they appear. The heat map shows whether these contributions are evenly distributed or focused on specific parts of the rankings.

yield a very similar set of domains, but may however not capture sudden increases or decreases in popularity.

Figure 4.2 shows that the default Tranco ranking has a much higher stability than the Alexa and Umbrella ranking, at around 0.6%, on par with Majestic and Quantcast. Figure 4.3 shows that this stability extends to smaller subsets of the ranking, with usually less than 1% daily change. Changes in the top 10 or 100 do occur, but are normally limited to one domain changing between consecutive days.

Figure 4.4 shows that aggregating over 1 day already removes a large part of the volatility introduced by the Alexa ranking. Moreover, for longer periods than the default 30 days,

the improvement in stability is relatively small. Finally, the 'weekend effect', where the set of domains in rankings based on weekend traffic differs significantly from those in weekday rankings [416], is largely subdued when aggregating over longer periods, including the default 30 days.

When a ranking is calculated for an aggregation period longer than 1 day, the set of component rankings will only differ in the first and last day of the period, meaning that only these two days affect the stability between the rankings of two consecutive days. To assess whether Tranco maintains stability over a longer period of time, we analyze the difference between rankings that are spaced apart with the length of their aggregation period. Figure 4.5 shows that rankings aggregated over 30 days or less produce a relatively stable set of domains, changing around 10% over time when no anomalies are present. Rankings aggregated over longer periods are less stable; moreover, anomalies (such as those discussed in Section 4.3.6) have a longer-lasting negative effect on stability.

Finally, Figure 4.6 shows the global change of the set of domains in Tranco over one year. For the full top million, 33.13% of domains are new when comparing the rankings of April 30, 2018 and April 29, 2019. For smaller subsets of the ranking, this global change is even lower: the top 10,000 sees only a 16.66% change. The observed level of global change is comparable to that of Majestic and lower than that of Alexa and Umbrella [427], and shows that Tranco provides stability even in the long term while still capturing genuine changes in popularity over time.

Existing rankings suffer from a high volatility, sometimes unknown to researchers [291, 427]. By aggregating over longer periods, Tranco provides a set of domains that does not change significantly over time, reducing the influence of the exact date on which the ranking is downloaded and therefore improving the soundness and reproducibility of (longitudinal) studies.

4.3.4 Responsiveness

If the websites of the domains in a ranking are unavailable, the size of the studied sample shrinks, which makes a study less comprehensive. Moreover, websites that produce an error when visited may not be representative of 'normal' websites and could therefore skew measurements.

Table 4.3 shows the distribution of HTTP status codes for a crawl of the root web pages for the Tranco ranking of May 14, 2019[9] conducted from May 14 to 16, 2019. For the full ranking, we find that 85.17% of domains respond with HTTP status code 200, indicating an available website. 4.14% report another status code, indicating that the server is responsive but that no content is provided on the root page of the domain. Finally, 10.68% of domains could not be crawled. Smaller subsets of the ranking see an increased success rate.

[9] Available at `https://tranco-list.eu/list/666X/1000000`.

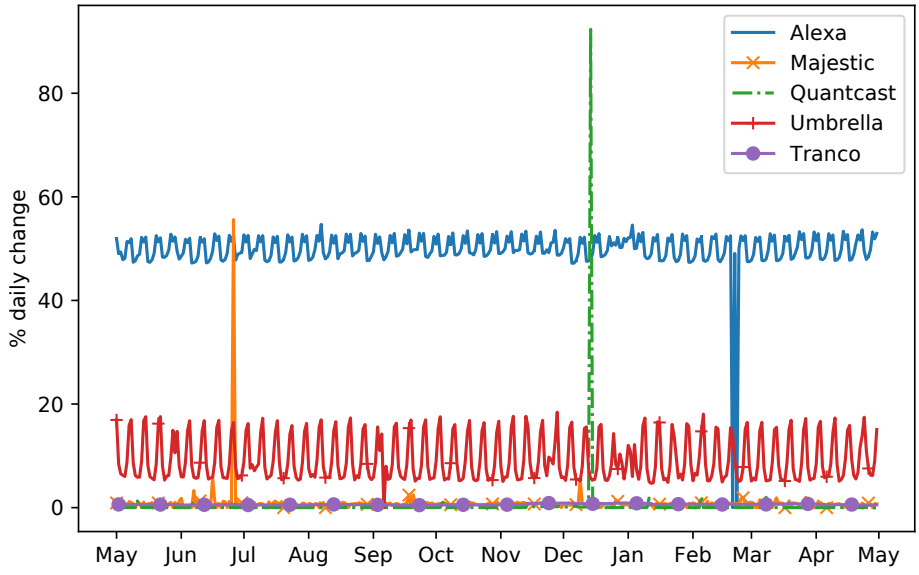


Figure 4.2: Stability over time from May 2018 until April 2019 of Tranco and its four component rankings, measured as the difference between rankings of two consecutive days.

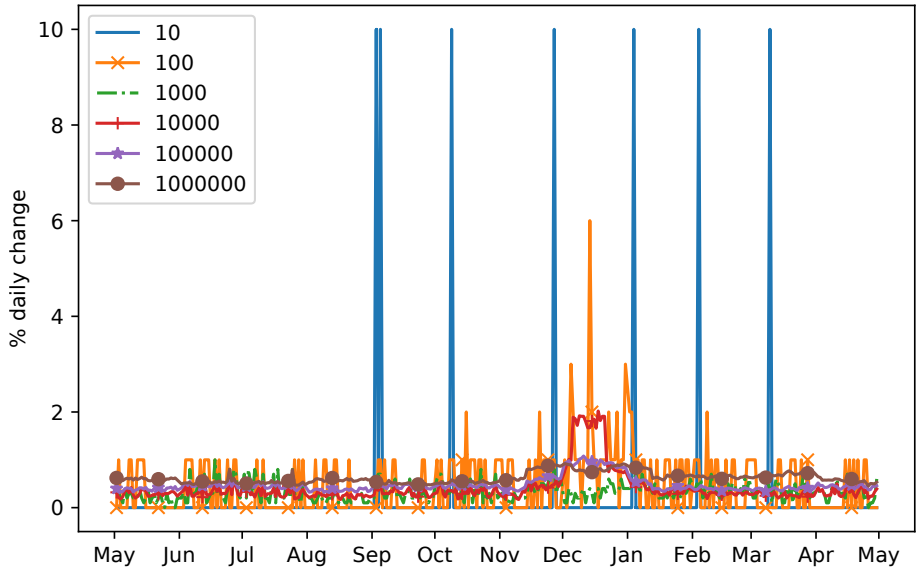


Figure 4.3: Stability over time from May 2018 until April 2019 of Tranco for different subsets, measured as the difference between rankings of two consecutive days.

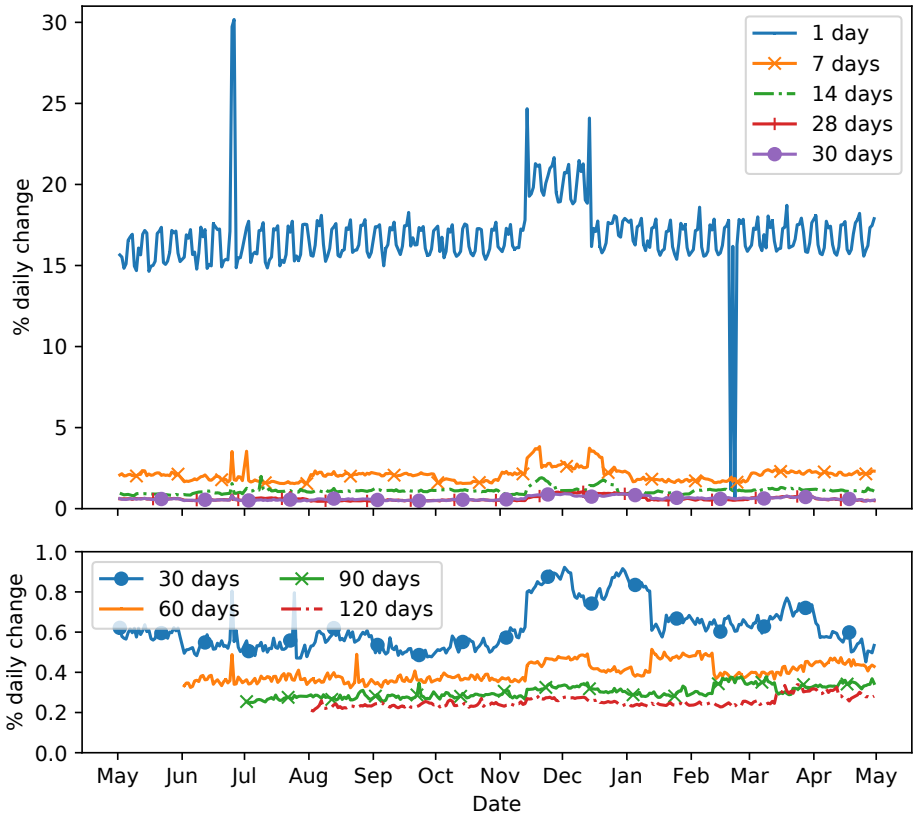


Figure 4.4: Stability over time from May 2018 until April 2019 of Tranco for eight aggregation periods, measured as the difference between rankings of two consecutive days.

In order to understand the source of crawl failures, we manually analyze the 109 domains in the top 1,000 that do not have status code 200. 32 domains are used for analytics or advertising, serving tracking scripts and advertisements or forming part of a redirection chain. 27 domains are used in a content delivery network, usually serving content that is embedded on other (popular) sites. 8 domains are part of some other type of infrastructure, e.g. serving Windows updates (`windowsupdate.com`) or as a portal for mobile users (`metropcs.mobi`). 22 domains blocked our crawler with e.g. a `403 Forbidden` status code. Finally, 20 domains failed to be crawled for some other reason, e.g. only serving content on a subdomain. This analysis shows that Tranco also covers domains that experience high traffic volumes but through other means than website visits. This complements e.g. Alexa's focus on website traffic and therefore makes Tranco appropriate for a broad set of studies (e.g. into the security of popular non-website domains).

By filtering out domains that do not appear in the Chrome User Experience Report

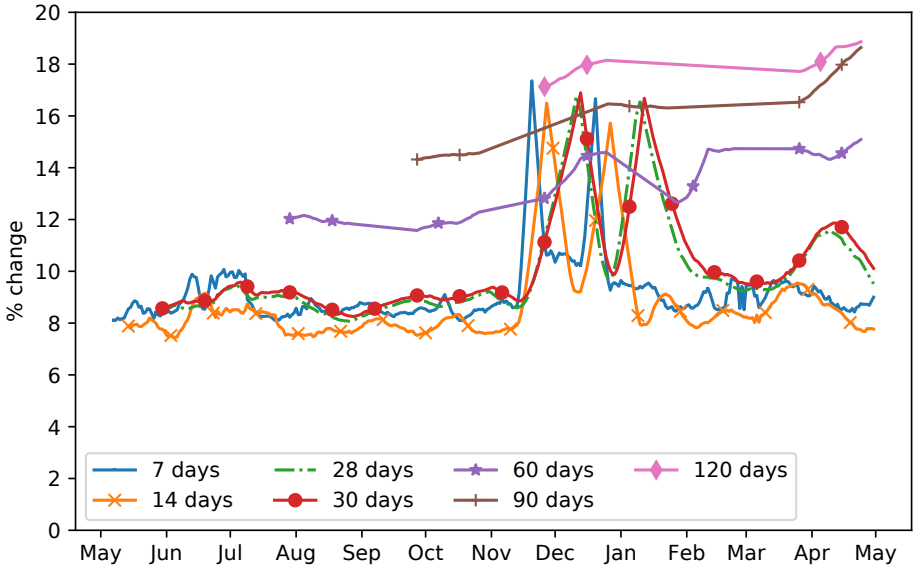


Figure 4.5: Stability over time from May 2018 until April 2019 of Tranco for eight aggregation periods of N days, measured as the difference between the rankings of dates $D - N$ and D.

Table 4.3: Responsiveness of (subsets of) the sites in the Tranco ranking of May 14, 2019.

Subset	Success (`200`)	Client error (`4xx`)	Server error (`5xx`)	Other status code	Failure
10	100.0%	0%	0%	0%	0%
100	92%	2%	0%	0%	6%
1k	89.1%	3.4%	0.2%	0.2%	7.1%
10k	86.24%	3.97%	0.57%	0.08%	9.14%
100k	84.99%	3.32%	0.80%	0.09%	10.79%
1M	85.17%	2.77%	1.30%	0.07%	10.68%

(Section 4.3.2), i.e. domains that are not regularly seen in a major browser, the share of responsive domains increases: 93.20% of domains from Tranco seen in the Chrome User Experience Report respond with status code 200, 1.89% respond with another status code, and 4.91% could not be crawled.

For the full Tranco ranking, Majestic shares the most sites that do not respond with status code 200, at 96,839 domains; Alexa shares 37,343 domains, Umbrella 34,486 and Quantcast 24,823. However, the full Majestic ranking does not have the most unresponsive domains of all four rankings [291]. The higher share of unresponsive domains contributed by the Majestic ranking may therefore rather be explained by the higher contribution of

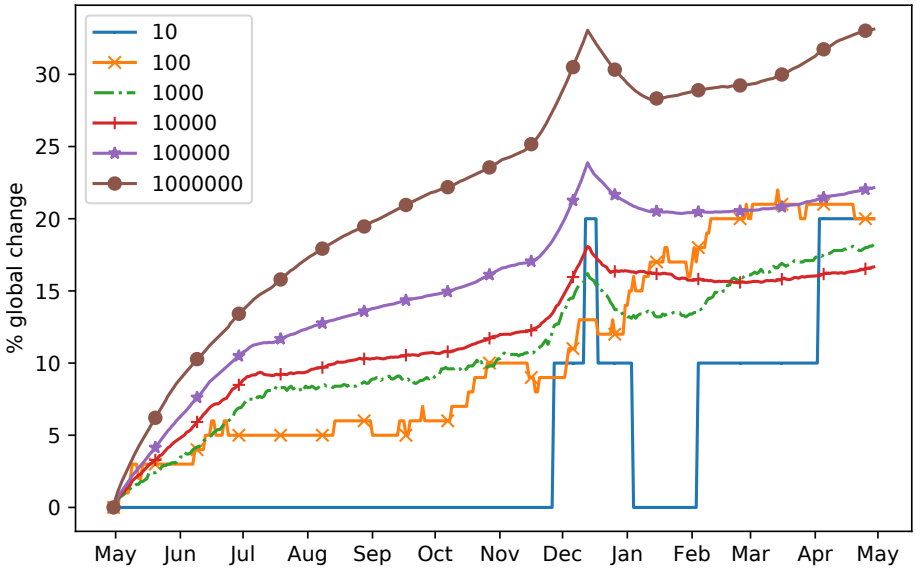


Figure 4.6: Difference between the Tranco ranking of April 30, 2018 and rankings from May 2018 until April 2019 for different subsets.

Majestic to Tranco overall.

Overall, the large majority of Tranco domains hosts a responsive website. This means that crawling those domains provides a sufficiently large sample, allowing for a broader and more representative overview of the web.

4.3.5 Benignness

Even though researchers and security companies often assume that very popular sites are by definition benign, and therefore whitelist them, malicious domains are present in top sites rankings [291]. Conversely, the rank of popular malicious domains could indicate how many victims they affect.

Table 4.4 shows that 1,851 unique domains on the Tranco ranking of May 14, 2019[7] were flagged by Google Safe Browsing [418]. 80.3% of these perform 'social engineering' (e.g. phishing). Within the top 10,000, four sites were flagged. Note that this represents a lower bound of malicious domains: unflagged domains can be benign but their malicious character could also not yet have been discovered.

Most malicious sites can be attributed to the Majestic ranking, at 1,168 domains; 703 malicious domains appear in the Alexa ranking, 594 in Quantcast and 210 in Umbrella. Of the four top 10,000 sites that are flagged, all appear in the Majestic ranking, 3 in Alexa and Umbrella, and 1 in Quantcast. Although the higher share of malicious domains from

Table 4.4: Number of sites in the Tranco ranking of May 14, 2019 that were flagged by Google Safe Browsing.

	10*k*	100*k*	1*M*
Malware	1	24	187
Social engineering	1	21	1,486
Unwanted software	2	34	189
Potentially harmful application	0	0	8
Total (unique domains)			1,851

the Majestic ranking may again reflect the higher contribution of Majestic to Tranco overall, Majestic does tend to include more malicious domains in and of itself [291].

Although popular websites are often assumed to be benign, we find that the four rankings that are aggregated into Tranco, and therefore also the Tranco ranking itself, still contain some malicious domains. Given that Google Safe Browsing can be reliably queried for one million domains, the Tranco ranking can be prefiltered to exclude the malicious domains.

4.3.6 Anomalies

Throughout the one year of rankings that we evaluate, we observe three major anomalies in the four component rankings. These have (temporary) effects on the composition of the Tranco ranking: they explain the peaks and sudden jumps observed in our results throughout this paper.

On June 25, 2018, the Majestic ranking was truncated, containing only 445,000 instead of the usual million domains. Between November 14 and December 13, 2018, Quantcast's ranking contained only around 38,000 instead of the usual 500,000 domains, discarding all domains for which traffic was estimated [291]. Finally, on February 20 and 22, 2019, the Alexa ranking was a duplicate of the previous day's list.

These anomalies mainly cause a higher or lower than average daily change in composition (as can be seen in Figure 4.4), due to changing contributions of the four component rankings. However, this effect is more outspoken for shorter aggregation periods, so larger periods including the default 30 days can smooth out these anomalies.

Moreover, we see that anomalies are not limited to one particular ranking, showing that real-world data collection and processing is susceptible to error. Tranco reduces the impact of these anomalies by aggregating data from multiple providers, such that sudden changes in one ranking's composition do not immediately result in a widely varying set of domains.

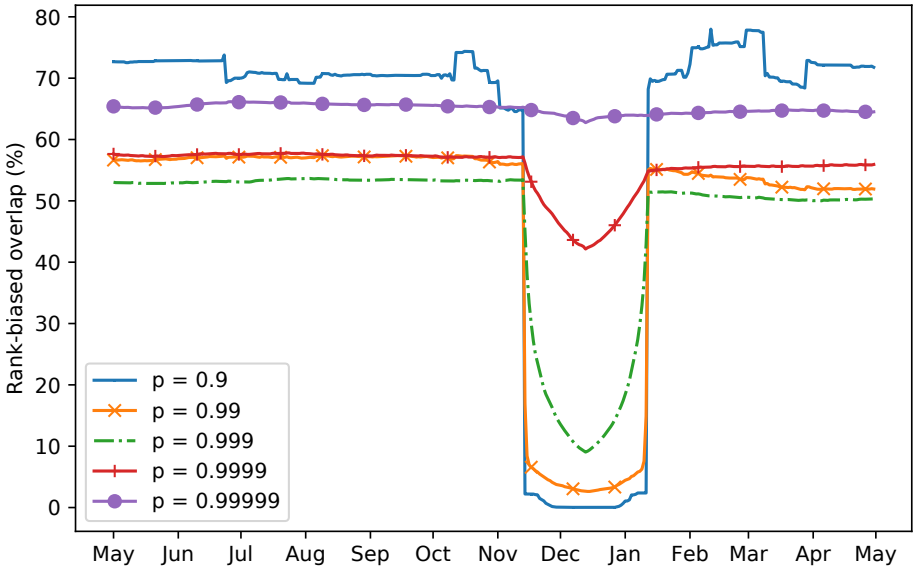


Figure 4.7: Similarity over time from May 2018 until April 2019 between rankings using the two scoring methods available in Tranco, measured as the rank-biased overlap between rankings of two consecutive days.

4.3.7 Combination method

The Tranco ranking supports two methods of calculating domain scores: Borda (total number of items minus rank) and Dowdall (inverse of rank) [185]. The latter reflects previous observations of a Zipf-like distribution in domain popularity [21, 121]. Figure 4.7 shows that these two methods produce moderately similar rankings, with a rank-biased overlap [513] of between 50% and 80% depending on its parameter p, which indicates how heavily a smaller subset of the ranking is weighted. In terms of stability, Figure 4.8 shows that the Borda method produces a slightly more stable ranking, but overall stability is comparably high.

However, Figure 4.9 shows that for small subsets (e.g. the top 1,000) the Dowdall method is more robust against anomalies in the data. This is due to the rescaling of ranks, which gave the anomalous Quantcast rankings in November and December 2018 a disproportionately high influence on the Tranco ranking. While the combination method produces similar sets of domains, the default Dowdall method results in a more stable list if anomalies are present.

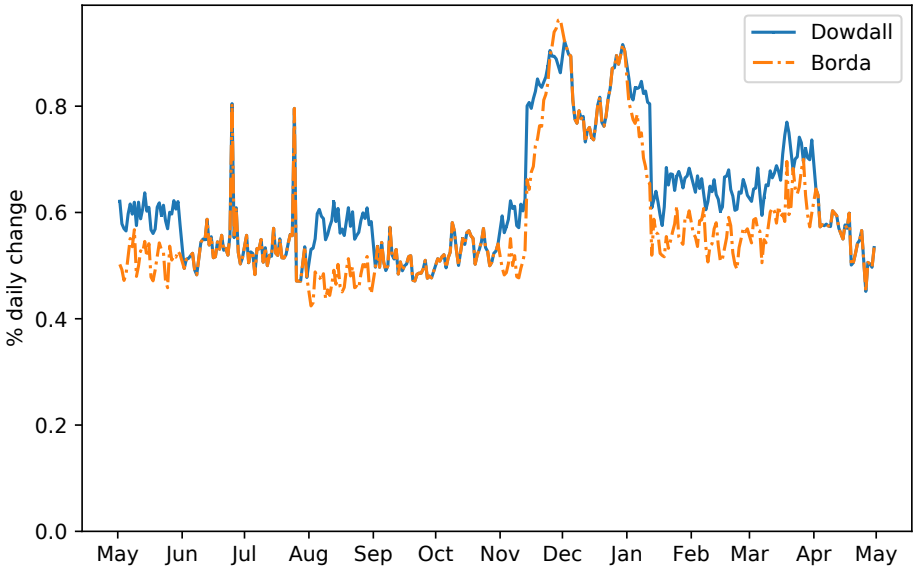


Figure 4.8: Stability over time from May 2018 until April 2019 between the two combination methods available in Tranco for the full ranking, measured as the difference between rankings of two consecutive days.

4.3.8 Structure

The default Tranco ranking only includes only 'pay-level domains', referring to domain names that a customer can directly register, as certain top-level domains do not allow direct registrations under the TLD (e.g. `.uk`, requiring to register under e.g. `.co.uk`). Due to the way in which pay-level domains are determined, by checking domains against the Public Suffix List [349] to extract the TLD and then taking the next label as the pay-level domain, no subdomains nor invalid domains remain in the default Tranco ranking. This means that the final set of domains captures a higher variety of valid hosts, content and ownership, allowing for more comprehensive studies.

4.4 Related work

Scheitle et al. [427] study the stability and similarity of the Alexa, Majestic and Umbrella lists, measure the potential impact on (Internet measurement) research and list guidelines for using the rankings in a sound and reproducible manner. Rweyemamu et al. [416] conduct a more detailed analysis of three effects of the rankings' methods on their composition, using these findings to extend the ranking usage guidelines.

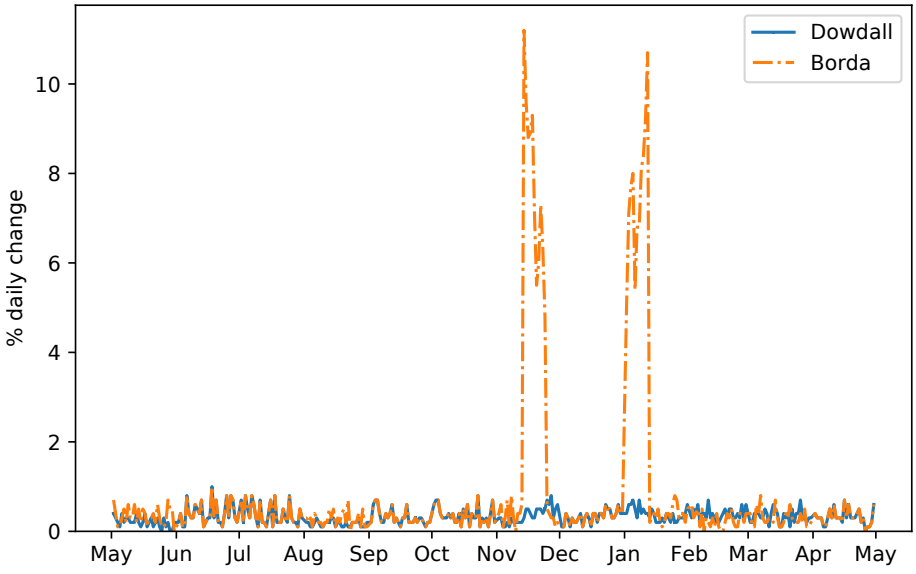


Figure 4.9: Stability over time from May 2018 until April 2019 between the two combination methods available in Tranco for the top 1,000, measured as the difference between rankings of two consecutive days.

In prior work [291], we analyze the Alexa, Majestic, Umbrella and Quantcast rankings within the context of security research: their analysis includes the representativeness, responsiveness and benignness of rankings. Moreover, we demonstrate that the rankings are susceptible to (malicious) large-scale manipulation. Finally, we provide Tranco, the ranking that we analyze in this paper, as a research-oriented, reproducible alternative to existing top websites rankings.

4.5 Conclusion

Many studies in web security and Internet measurement research depend on rankings of popular domains. We presented the Tranco ranking as a more research-oriented alternative to existing rankings, emphasizing a transparent method and providing a publicly available archive of reproducible rankings. However, the influence of its parameters on the composition of the ranking had not yet been analyzed. We evaluate Tranco over one year, and find that it has the following research-related properties:

Agreement with existing rankings The more stable Majestic and Quantcast rankings share the most domains with Tranco. Alexa shares fewer domains due to its volatility, but these domains are highly ranked. Umbrella contributes the least to Tranco, as only 'pay-level domains' are listed in Tranco by default.

62% of domains in Tranco were observed to be regularly visited in the Chrome browser, indicating that these are genuinely popular websites.

Stability The default aggregation period yields a very stable ranking; with a smaller period, some volatility is reintroduced, while a larger period only causes minor gains in stability.

Responsiveness When crawling their root page, Tranco contains around 15% unresponsive domains. However, many of these domains do play an important infrastructural role, showing how Tranco captures a balanced set of popular domains beyond regular website traffic.

Benignness Tranco contains around 0.2% domains flagged as malicious.

Anomalies The default Dowdall scoring method is more robust against observed anomalies in Tranco's component rankings.

Structure Tranco contains only valid pay-level domains.

Subsets In general, overlap with existing domain lists, stability and responsiveness improve slightly for smaller subsets of Tranco.

Moreover, these properties can be further improved by creating a customized ranking where appropriate filters are applied.

Our analysis informs researchers who need to use a top websites ranking on those characteristics of Tranco that might be important to their study, and serves as a guide for those who want to customize the ranking to their requirements. In particular, we find that the default parameters selected for the daily updated Tranco ranking, available at `https://tranco-list.eu/`, are overall appropriate and recommended for research such as large-scale security measurements that needs a representative set of domains.

Part II

Auditing automated decision-making systems

5
Prologue

In the next two chapters, we present two case studies where we critically evaluate *automated decision-making systems*, which use algorithms and automated models to process decisions at scale [341]. While these systems are beneficial for coping with an ever-increasing quantity of decisions that quickly become unfeasible to process manually, there is a risk of erroneous decisions, disadvantageous and discriminatory biases, or opaque decision-making processes, all posing ethical challenges to deploying such systems [138, 187, 216, 341, 466].

The success of an automated decision-making system depends on several components. The system needs data for training its models and making a decision on new instances. This data needs to be representative as to have the models learn the correct concepts for making accurate decisions. Biases within this data may be exacerbated by the models and propagate to the system's outputs, risking false or discriminatory decisions if the data is not properly selected [138, 216, 466]. The composition of the data set can also affect the accuracy of its results: for example, in detecting political ads, selecting different sources for training data yield different models with varying performance [456]. The data also needs to be well available and of sufficient quality for the model training to be useful and sound. The models themselves need to be designed appropriately according to the data that it needs to work with. These models can range from simple rule-based systems to deep neural networks [341], but all must be properly parameterized, evaluated, and deployed for them to be functioning correctly. Finally, there may be different requirements for the outputs of the models, which in turn restrict which models can be used. For example, certain use cases or even legal instruments may require that the model generates an explanation of how the decision was made [20, 338, 503], or an indication of its confidence in its decision. In general, the outputs must be usable and complete as to not have instances where no or an uncertain decision is made. Bluntly put, a variety of failures can cause an automated decision-making system to simply not work [397], and all components of such a system must be properly set up and balanced to achieve any success.

Automated decision-making systems are also finding their way into security systems,

again to discover security-related events within the vast amounts of data that passes through these systems. Automated decision making is used for applications such as detecting spam [214] or phishing [45], intrusion detection [336], discovering malware binaries [366] and domains [55], or authentication [488]. However, the effectiveness of deploying such systems in practice has been questioned, and major challenges related to data, models, and outputs have been identified. For example, machine learning-based solutions for intrusion detection may not be very effective [452], and the lab performance of machine learning models for malware detection cannot be reproduced in real-world settings, making them unreliable and untrustworthy [108]. This low effectiveness was found to be the primary reason why security practitioners are reluctant to adopt machine learning tools [334]. In security research, studies commonly fall victim to pitfalls in the design, implementation, and evaluation of machine learning-based solutions for security problems, leading them to overestimate the achieved results, and ultimately harming the validity and soundness of these studies [59]. For example, (reported) metrics used for evaluating machine learning-based research proposals for authentication systems may be flawed [465].

Ethical concerns about biases and unfairness in automated decision-making systems have given rise to algorithm audits [76, 127, 521], which are designed to broadly evaluate whether such a system meets its performance claims without harmful effects [143]. The intention is to improve algorithmic accountability [127, 143], or the practice of holding entities that deploy automated systems accountable and responsible for ensuring these systems work properly and fairly without harms. Audits vary by who conducts them – an internal team, a contractor, or an independent third party [127], and the way in which data is collected – e.g., scraping, test accounts, or crowdsourcing [420], but usually consider the decision process as an opaque system where only the inputs and outputs are visible [278, 420]. Algorithm audits go beyond simple transparency: just transparently publishing artifacts such as the source code of an algorithm or the machine learning method used is insufficient for observing whether the decision process itself works fairly, and is therefore insufficient for full accountability [265, 278]. However, transparency remains a crucial enabler for allowing audits to happen, of all components including of the data that drives the algorithms [143, 145, 155, 156], although this must be weighed against concerns about privacy, commercial interests, and manipulation [127, 143, 278].

In our two case studies, we incorporate these angles on critically assessing the worth of automated decision-making systems in security applications, by evaluating the merit of a "human-in-the-loop" approach to reduce the risk of errors, and conducting an audit to independently evaluate the performance of a decision-making algorithm.

Avalanche botnet takedown In the first case study, we propose an approach using an automated model to assist law enforcement investigators during a botnet takedown in resolving collisions of benign domains with presumably malicious domains generated and used by that botnet. Next to designing a generic approach, we evaluate it in a real-world context. Here, this is the takedown of the Avalanche botnet, where law enforcement sought to prevent attackers from maintaining or gaining control of the

domain names that the machines in the botnet use to communicate with the command and control servers. These domains are generated pseudorandomly using Domain Generation Algorithms (DGAs), but coincidental collisions with legitimate domains are possible. These legitimate domains must first be identified to avoid that they become part of the domain takedown. Our use case is orthogonal to that of prior work, as we essentially detect benign domains within a set of mostly malicious domains, instead of previous attention to detecting malicious domains in otherwise benign traffic. This different use case warrants a custom approach that requires different features for the automated decision-making model. As another example of such a diverging use case, COMAR [318] considers the problem of discerning maliciously registered from compromised domains. Both are malicious in behavior, but the registration intent differs, requiring different takedown actions. Maliciously registered domains can safely be taken down to prevent any further abuse, but since compromised domains also host benign content, mitigation should focus on selectively targeting and removing the malicious content while not harming the availability of the benign content. Automated decision-making models must therefore be adapted to and evaluated for each use case separately, as their decisions may have significantly varying consequences and may become harmful if not made and applied correctly.

Our approach supports this component of the botnet takedown, by providing a machine learning model that automatically classifies domains generated by Avalanche DGAs as benign or malicious. We go beyond the pure development of an automated model, and enrich both the approach and evaluation to improve the reliability for law enforcement investigations. First, our approach leverages a synergy of automated and manual decision-making, overcoming the shortcomings of an automated-only approach. Concretely, we start from an automated prediction, but use a model that outputs a confidence score for its prediction. The investigator can then set a threshold for tolerated error rates, which translate into bounds on the confidence scores. Domains with scores within these bounds, i.e., domains for which the model is too uncertain, then have to be verified manually. Optionally, two separate thresholds could be selected separately to incorporate the different risks attached to the two error types: false positives, which cause benign domains to be taken down unjustly, and false negatives, which leave domains available for abuse and potential respawning of the botnet. The error tolerance then reflects the trade-off between classification accuracy and manual effort reduction. For a lower error tolerance, the model will be considered uncertain for more domains, which means more domains have to be classified manually, but with the benefit of greater accuracy. However, manual investigation effort can still be significantly reduced while maintaining high accuracy: for example, applying this approach to the 2019 iteration of the Avalanche takedown results in a reduction of 76.9% in manual effort. The human-in-the-loop approach [60] that we designed therefore strikes a realistic balance that acknowledges both the ability for automated decision-making systems to contribute to classification at scale, and the risk from potential errors that any such automated decision-making system may make, an insight that can also be helpful for future takedowns, or other security applications. For example, Miller et al. [332] already showed that such an approach also improves accuracy in labeling malware binaries, and Aonzo et al. [57] found that humans

and machines classify malware binaries using different features, meaning they can work complementarily.

Second, we develop and evaluate our approach under a real-world setting where certain desirable preconditions are not fulfilled, as opposed to a theoretical exercise with an optimal setting. This means that we must show that our decision-making system works in a more challenging but also more realistic environment. On the one hand, the Avalanche takedown has three unique characteristics that impose constraints on our approach and prevent us from relying on existing approaches: bulk patterns that were previously used to determine maliciousness are no longer present, the classification must be done proactively before malicious activity is observable, and no active connections to domains can be made as to not reveal the ongoing investigation. We overcome these constraints by developing a rich feature set that represents only individual, proactive, and passively obtained patterns, but must therefore also omit certain potentially useful features that violate one or more constraints. On the other hand, our ground-truth data set is inherently small, as collisions between benign and DGA-generated domains are relatively rare. Moreover, the data sets required for our feature set do not always have data for all domains in our ground truth. These are challenges that law enforcement and by extension any designer of an automated decision-making system would or could also face. We adapt our design and evaluation to account for these restrictions, which make them valuable in analyzing the feasibility of an automated approach for their use case. In our approach, we account for missing data through an "ensemble model", where we train a model for each combination of available data sets. Moreover, given that not all data sources are equally easy to collect, we evaluate their impact on the correctness of our classification. Our results show that unavailable data sets make our decision-making system less effective, but we can still achieve reasonable performance when data is missing.

Using the model that we developed, we assisted law enforcement in the 2019 iteration of the Avalanche takedown, which sought to prevent the abuse of DGA domains for the next five years. Even though some erroneous takedowns occurred [91, 371], overall, the takedown appears to have been effective at continuing to disrupt the Avalanche botnet. Still, the registration blocks and domain seizures that resulted from the takedown must be maintained, and the sinkhole infrastructure supporting these measures remain operational to this day. As millions of machines remain infected [66], any lapse could allow an attacker to register a domain contacted by these machines, and therefore could allow the botnet to respawn.

Facebook political ads In the second case study, we audit the large-scale automated decision-making system deployed by Facebook for detecting ads that violate its policies on political advertising. This system can be considered a prime example of a model deployed at a massive scale by one of the largest technology platforms, who should be able to invest many resources (technical, financial, human, and otherwise). Social networks are among the largest platforms that face the challenge of processing the high volume of content that is generated on or passes through their platform, in particular as the content is usually user-generated. As such, they deploy automated models that

make decisions for actions such as content moderation, including policy enforcement, or recommendation engines.

Prior work has audited large platforms on the effects of using automated systems for their advertising practices as well as potential influences on democracy. Facebook's automated systems for ad delivery have been studied for biases in the ads that users see, potentially leading to discrimination. Lambrecht and Tucker [283] found Facebook's algorithms to introduce a gender bias into the delivery of STEM career ads. Ali et al. [42] found Facebook's ad delivery to be influenced by an ad's budget, content, and image, leading to skewed delivery based on, a.o., gender or race, including for job and housing ads. They later found such biases to also exist for political ads [43], as Facebook tended to show an ad to users whose inferred political preference agreed with the ad's message. Imana et al. [241] developed a method for auditing Facebook's ad delivery for job ads, and found it to be skewed by gender. Papakyriakopoulos et al. [374] measured how algorithms affect the distribution of political advertising across Facebook, Google, and TikTok. They found that the distribution of ad impressions and prices across demographic groups is skewed compared to the platform's demographic distribution, and that moderation decisions for the same or similar ads are inconsistent within the platforms. Sapiezynski et al. [422] found that 'Lookalike Audiences' and 'Special Ad Audiences', both tools that allow advertisers to generate ad audiences similar to a list of users that the advertiser provides, exhibit similar biases in terms of the demographics of selected users, even though the algorithm for Special Ad Audiences is designed to specifically exclude these demographic parameters. Their work is a case study for showing that, given the complexity of algorithmic systems, simply removing features may not be sufficient for reducing discriminatory outcomes.

YouTube's recommendation algorithms have been particularly audited for whether they lead users to (more) content related to misinformation. Hussein et al. [238] found little evidence for demographic characteristics to immediately affect whether new users see misinformation videos on YouTube. However, once users built a watch history, the recommendation algorithm suggested more misinformation content to them, keeping them in a 'filter bubble' [375]. Ribeiro et al. [402] found that YouTube recommended that users within communities linked to radicalization watch channels and videos within their own community, but also other radicalization-related communities. Spinelli and Crovella [458] found that YouTube's recommendations tend to lead users away from reliable sources. Haroon et al. [223] found that YouTube's algorithm generates ideologically biased recommendations, which are especially present for right-leaning users. They propose interventions to mitigate this bias, such as essentially 'confusing' YouTube's recommendation algorithm by playing bias-reducing videos in the background. Papadamou et al. [372] found that watch history has a significant effect on search results and video recommendations on YouTube in the space of pseudoscientific content, again contributing to a personalized 'bubble' that users may find hard to escape. Tomlein et al. [475] searched ways to 'escape' this bubble created by YouTube's recommendations, including watching videos debunking misinformation, which they found to be generally effective. Faddoul et al. [171] found that policy changes by YouTube have reduced the number of recommended conspiracy videos, although the filter bubble effect has not yet

fully disappeared.

More broadly, Narayanan and Lee posited the need for independent security policy audits [355]: understanding problems and flaws in developing, implementing, and enforcing security-related policies and processes. In their research agenda, they already highlight audits of platforms as a research direction, on topics such as content moderation, advertising, and algorithms. In a similar spirit, Simko et al. [441] make a case to conduct regular and automated audits, specifically for social media platforms and the impact of their recommendation engines on misinformation spreading. The goal is to improve the accountability of these platforms, complementing their self-regulatory policies with independent assessments of the platforms. Ali [41] presents a research agenda for measuring and mitigating biases in online advertising introduced by recommendation algorithms, specifically focusing on involving users to understand their perceptions of biases and harms. Matias et al. [321] propose a framework for software-supported audits, and discuss the design considerations to conduct these audits as realistically, statistically correctly, and ethically as possible. Imana et al. [242] propose a framework for platform-supported audits, and describe the requirements to conduct such audits in a privacy-preserving manner. Our independent audit falls within these research agendas, serving as a gauge of the current capabilities of advanced automated decision-making systems at scale, for one highly scrutinized decision system on the web today – political ad detection. Within the audit framework of Costanza-Chock et al. [127], our audit is a third-party audit, as we are "independent researchers with no contractual relationship to the audit target" [127]. Within the audit framework of Sandvig et al. [420], our audit is a scraping audit, where we retrieve our data set through "repeated queries to a platform" [420].

Since our work, Facebook expanded the Ad Library and the authorization requirement to nearly all countries [64], compared to only around half at the time of our work. This equalizes transparency worldwide and improves fairness on the account of the benefits that this brings to the electoral process. Within the Ad Library, the most significant addition in mid 2022 was aggregated metadata about which audiences an advertiser targeted [266]. This metadata contains the ad count and spend share for political ads for each type of targeting: location, age, gender, interests, language, custom audiences, and lookalike audiences. However, its aggregated nature limits the insights that can be obtained into ad targeting by political advertisers, as for example certain discriminatory practices could be lost in the aggregate. Per-ad metadata is available only to vetted researchers registered to the Facebook Open Research and Transparency project [169]. This project has been criticized for its time-limited data set, closed environment for data analysis, and the requirement that Facebook reviews papers before publication [158, 373]. Moreover, the availability of this additional metadata suffers from the same fundamental shortcoming as identified in our work: by restricting it only to ads known by Facebook to be political, it is crucially missing for those political ads that Facebook fails to detect. On the front of data access, the web portal now provides downloadable CSV files, improving automated processing of active ads metadata in particular. However, these files cannot be requested or retrieved automatically, and one user is limited to 3 CSV exports per day. The CSV export is therefore not a replacement for the custom automated data collection of active ads that we developed for our study.

In aggregate, the restrictive posture of Facebook towards (independent) researchers suggests that the poor performance of Facebook's enforcement systems that we find in our audit may not be due only to the technical limitations in deploying automated decision-making systems. Instead, the performance of these systems also depends on the platform's priorities in terms of the infrastructural, human (reviewer and management), up to financial resources invested to address the issue of political ad detection and content moderation at large. The economic incentives for platforms like Facebook may also play a role here: the fact that in the process of removing (violating) ads, the platform may lose advertiser revenue, creates an inherent tension with their larger business goals; that said, Facebook has stated that it "views its political-ad business as a civic responsibility rather than a revenue driver" [201]. External actors may need to intervene to apply corrections to these incentives in order to encourage and require platforms to improve (the automated solutions used for) policy enforcement. In this direction, legislators in both Europe and the US have filed proposals to regulate online political ads, in part due to upcoming major elections in 2024 for both. In the EU, a proposed regulation addresses "transparency and targeting of political advertising" [393], with the Digital Services Act also addressing online advertising in general, with particularly strict requirements for very large online platforms such as Facebook [392]. In the US, the proposed Platform Accountability and Transparency Act is specifically meant to support independent research into large online platforms by requiring them to provide access to data [113]. The new EU legislation allegedly prompted Meta to consider scrapping political ads in Europe altogether [166], in part precisely because of Meta's perception that enforcing new regulations correctly will be difficult. These developments highlight that the debate on how online political advertising should be approached, which restrictions should be imposed upon it, or even whether it should be allowed altogether, is still very much ongoing.

6

A Practical Approach for Taking Down Avalanche Botnets Under Real-World Constraints

This chapter is based on the homonymous paper published in the proceedings of the 27th Annual Network and Distributed System Security Symposium (NDSS 2020) [292]. This work was co-authored with Tim Van hamme, Sourena Maroofi, Tom Van Goethem, Davy Preuveneers, Andrzej Duda, Wouter Joosen, and Maciej Korczyński.

In 2016, law enforcement dismantled the infrastructure of the Avalanche bulletproof hosting service, the largest takedown of a cybercrime operation so far. The malware families supported by Avalanche use Domain Generation Algorithms (DGAs) to generate random domain names for controlling their botnets. The takedown proactively targets these presumably malicious domains; however, as coincidental collisions with legitimate domains are possible, investigators must first classify domains to prevent undesirable harm to website owners and botnet victims.

The constraints of this real-world takedown (proactive decisions without access to malware activity, no bulk patterns and no active connections) mean that approaches from the state of the art cannot be applied. The problem of classifying thousands of registered DGA domain names therefore required an extensive, painstaking manual effort by law enforcement investigators. To significantly reduce this effort without compromising correctness, we develop a model that automates the classification. Through a synergetic approach, we achieve an accuracy of 97.6% with ground truth from the 2017 and 2018 Avalanche takedowns; for the 2019 takedown, this translates into a reduction of 76.9% in manual investigation effort. Furthermore, we interpret the model to provide investigators with insights into how benign and malicious domains differ in behavior, which features

and data sources are most important, and how the model can be applied according to the practical requirements of a real-world takedown.

6.1 Introduction

On November 30, 2016, a global consortium of law enforcement agencies and Internet stakeholders completed a four-year investigation aimed at dismantling the Avalanche infrastructure [1], which has been called "the world's largest and most sophisticated cybercriminal syndicate law enforcement has encountered" [505]. For seven years, this 'bulletproof hosting service' [47] offered services to cybercriminal operations through a 'crime-as-a-service' model [505], fully managing all technical aspects of carrying out malware attacks, phishing, and spam campaigns. It supported a botnet of a massive scale: Avalanche was responsible for two thirds of all phishing attacks in the second half of 2009 [7], and ultimately affected victims in over 180 countries with estimations of its monetary impact reaching hundreds of millions of euros worldwide [369]. The takedown operation in 2016 was supported by authorities from 30 countries and culminated in five arrests, 260 servers being taken offline and the suspension of over 800,000 domains [1].

As part of this dismantling, a large domain takedown effort sought to disable the botnet's communication infrastructure. This effort targets the large sets of domains that the malware families of Avalanche generate through *domain generation algorithms* (DGAs). Through this 'domain fluxing' [388], infected hosts attempt to contact all generated domains, whereas the botnet master only needs to register one to continue operating the malware, decreasing the likelihood of blacklisting and takedown. However, as security researchers have reverse-engineered several of these DGAs [388], law enforcement is able to identify upfront which domains the malware will try, after which these can be blocked or seized. Over four yearly iterations of the Avalanche takedown, more than 4.3 million domains were thus prevented from being abused, making it the largest domain takedown so far [65].

Previous work related to DGAs focused on detecting *malicious* domains in regular traffic, relying on strong indicators of *ongoing* malware activity, to discover new malware families or find infected hosts inside a network [55, 447, 526]. In this paper, we address the orthogonal issue that the Avalanche takedown faces: given – presumably malicious – DGA domains that will be generated in the future and should *proactively* be taken down, we seek to detect those that accidentally collide with *benign* domains. In particular, we assess how we can effectively support law enforcement investigators with an automated domain classification to inform the appropriate takedown action in a real-world use case. This reduces the extensive manual effort previously invested in this classification, while still maintaining the high accuracy required in such a sensitive operation. Taking down benign domains may cause prejudiced service interruption and harm their owners. At the same time, we have to guarantee that no malicious domain is left untouched, as this would allow malicious actors to target infected users once again.

We are the first to develop an approach that can be used to effectively identify the domains registered with malicious intent, within the constraints of a real-world takedown operation. First, *bulk patterns* no longer apply, both for domains that are benign (due to the accidental uncoordinated collisions) and malicious (due to the low number of required domains). Second, as the takedown is *proactive*, we cannot search for malicious activity (any ongoing activity would mean that infected machines are implicated in actual attacks and defeat the proactive purpose of the takedown). Third, we *cannot actively contact domains* so that the takedown can occur stealthily (otherwise attackers could evade detection and undermine the takedown). Instead, we rely on capturing more generic differences in how benign and DGA-generated malicious domains are registered and operated.

We design a machine learning-based model for classifying benign and malicious domains, and we evaluate it on ground truth from the 2017 and 2018 iterations. Using a human-in-the-loop approach that combines automated classification and manual investigation targeted at the most difficult domains, we achieve an accuracy of 97.6% for the real-world Avalanche use case, ensuring high correctness while still vastly reducing manual effort: in the 2019 iteration, our approach reduced this effort by 76.9%. However, we go beyond reporting this metric with an extensive analysis of the benefits and limitations brought by the machine learning approach as well as the real-world setting. We provide an interpretation for the factors that impact the decisions of the model, giving insight into how the owners of benign and malicious domains behave differently and how the model uses this information to make decisions. These insights can help law enforcement in their choices regarding the acceptable performance and reliability of the model.

Malware creators increasingly employ techniques that make the takedown of their command and control infrastructure more complex, and the scale of malicious operations continually increases. Further automation of the takedown process with our classifier of malicious and benign domains can support law enforcement in coping with the increased complexity. However, we need to carefully design, evaluate, and analyze such an approach to cope with the constraints of a real-world application as to avoid any adverse effect on the legitimacy of the operation. This enables law enforcement to continue disrupting malware infrastructure and protecting potential victims.

In summary, our contributions are the following:

- We assess to what extent an automated approach can assist law enforcement investigators in correctly detecting the collisions with benign domains among registered domains implicated in the Avalanche takedown, without the ability to rely on bulk malicious registrations, ongoing malware activity or actively collected traffic.
- We develop a technique where we complement a machine learning model with targeted manual labeling of the most informative and difficult domains, to maintain performance across multiple takedown iterations while still vastly reducing the required manual investigative effort.

- We evaluate how well this approach performs and transfers for the 2017 and 2018 takedowns: we obtain an accuracy of 97.6%. The predictions of our model were used in the 2019 takedown, and we find a subsequent reduction in manual investigative effort of 76.9%.
- We critically examine the factors that impact the performance and decision-making process of our model. We find that time-based features are the most important ones, which at the same time are the most costly to evade. In terms of data set availability, WHOIS data greatly improves accuracy, which shows its importance for conducting effective cybercrime investigations.

6.2 Background

6.2.1 Domain generation algorithms

Machines in a botnet such as Avalanche communicate with the malicious actor through command and control (C&C) servers. Early malware hard coded the domain names or IP addresses of their C&C servers, so it was easy to obtain this information and either blacklist the servers or even take over the corresponding infrastructure (by pointing for instance the domains to 'safe' IP addresses and/or having hosting providers take C&C servers down), effectively stopping the malware from further malicious operation [78]. Malware has therefore evolved from hard coding the C&C server information to dynamically creating or updating it.

One technique of this dynamic approach is 'domain fluxing', in which domain generation algorithms (DGAs) create up to thousands of algorithmically generated domains (AGDs) every day [388]. The malware will then attempt to contact these domains and ignore the unavailable ones: the botnet owner therefore only needs to set up one of the generated domains to host a C&C server [78]. Avalanche combined this technique with 'fast fluxing', in which compromised machines hosting a proxy to the C&C server as well as the corresponding DNS entries of the AGDs rapidly switch [230], thus further evading blacklisting and takedown [1].

DGAs take as seeds parameters known to both the malware owner and the infected host, so that they both generate the same set of domains [78, 388]. These parameters such as the length of domains, top-level domains (TLDs) to use, or seeds for pseudo random number generators can be hard coded. More complex algorithms may depend on time: one of the inputs to the DGA is then the current time, either from the system clock or retrieved from a common source (e.g., GET requests to legitimate sites [527]). In this way, the DGA creates domains having a certain *validity period*: the time frame during which the seed timestamps make the DGA generate that domain, which the infected machines then attempt to reach. For Avalanche malware families, these validity periods range from 1 day (e.g. Nymaim) to indefinitely (e.g. Tiny Banker).

Table 6.1: Examples of domains generated by Avalanche DGAs.

	Domain	Malware	Validity
1	0a85rcbe2wb5n5fkni4i4y[.]com	CoreBot	Jan 21, 2018
2	researchmadness[.]com	Matsnu	Jan 28-31, 2018
3	arbres[.]com	Nymaim	Mar 9, 2018
4	sixt[.]com	Nymaim	always

We can further distinguish between deterministic DGAs that know all parameters upfront and non-deterministic DGAs that know some parameters only at the time of generating the domains: e.g., the DGA of the Bedep family uses exchange rates as seeds [433]. Avalanche did not use any non-deterministic DGAs so for successfully reverse-engineered DGAs [140, 388], we can generate all potential AGDs ahead of their validity, by varying the timestamp that serves as input to the DGA.

Table 6.1 lists example names generated by DGAs, from malware hosted by Avalanche. While Example 1 appears random (a long name with many digits and no discernible words), certain DGAs generate names that look much more like legitimate domains. Example 2 shows a name generated based on a word list yielding domains that may correspond to a regular domain name. Example 3 shows a short yet randomly generated name for which there is a high probability of generating either a valid word or a plausible abbreviation. These last two examples have a high probability of generating domains that collide with existing benign domains.

Finally, certain malware families alter domain resolution on the infected host, generating traffic to hard-coded and otherwise benign domains that actually resolve to malicious IP addresses to circumvent domain-based filters [224]. While these domains are not algorithmically generated, they are present in malware code and traffic and must therefore also be classified as part of the takedown operation, to distinguish them from other hard-coded and actually malicious domains. Example 4 is one such instance using the domain of the Sixt car rental site. We include these domains in our classification, but for brevity, we refer to all domains to be classified as the 'registered DGA domains'.

6.2.2 Taking down the Avalanche infrastructure

The perpetrators behind the Avalanche infrastructure offered two services for rent by cyber criminals: registering domain names as well as hosting a layered network of proxy servers through which malware actors could control infected hosts and exfiltrate stolen data [140]. Avalanche thereby supported the operation of 21 malware families [67], controlling a botnet of an estimated one million machines at the time of takedown [140].

Prosecutors completed the first iteration of the takedown in November 2016, where the whole infrastructure was dismantled through arrests, server seizures, and domain name

takedowns [1]. For the latter, the first iteration targeted live C&C domains, but also those that would be generated by the DGAs in the coming year, preemptively blocking these to prevent Avalanche from respawning. This effort has been repeated every year since, as in January 2020 infected machines on over two million IPs still contacted the Avalanche network [66], highlighting the potential damage if Avalanche were to respawn.

Coupled with the large number of malware families and the extensive amount of domains that these DGAs generate, this results in a large number of DGA domains to be processed. For the three yearly iterations from 2016 to 2018, this amounts to around 850,000 domains per year [65, 67], while the 2019 iteration looks ahead five years and therefore treats almost 2 million domains: this means more than 4.3 million targeted domains have been processed in total. For the DGA domains in the Avalanche takedown, law enforcement took one of three actions on the takedown date [391]:

- *Block registration:* for a not yet registered domain, the TLD registry blocks registration. This is the case for the vast majority of domains.
- *Seize domain:* for a domain registered by a seemingly malicious actor, it is seized from the original owner and 'sinkholed', i.e. it is redirected to servers of the Shadowserver Foundation. Optionally, domains are also transferred to the "Registrar of Last Resort". Through sinkholing, law enforcement can then track how many and which infected hosts attempt to contact the domains [66] and aid in mitigation through notifications to network operators and infected users [111]. Domain *seizures* require a legal procedure such as a court order, while organizations could also *request* a takedown through a 'takedown notice' [239].
- *No action:* for a domain registered by a seemingly benign actor (including domains sinkholed by other security organizations), no action is taken by law enforcement and the domain remains with its original owner.

6.3 Problem statement

6.3.1 Making accurate takedown decisions

The aim of the Avalanche takedown is to prevent the botnet owners from interacting with infected machines by blocking access to the required domains that the DGAs will generate in the year following the takedown. However, as these DGAs may generate labels that collide with benign sites, performing a blanket takedown of all generated domains would harm legitimate websites. For Avalanche, public prosecutors therefore first had to manually classify domains into benign and malicious: as shown in Table 6.2, they had to determine an appropriate action for a few thousand registered DGA domains each year.

For registered domains, an incorrect decision may have unintended adverse effects [117, 239]. In case of the seizure of a benign domain, its legitimate owner can no longer provide

Table 6.2: Number of benign and malicious domains per iteration.
*: according to our classification.

	2017	2018	2019–2024*
Benign	1397	1014	4945
Malicious	1145	402	1053
Classified	2542	1416	5998
Sinkholed	1177	594	2293
Total	3719	2010	8291

its service to end users. Owners may experience lengthy downtime, as challenging an illegitimate seizure and regaining the domain can be an opaque and difficult process [239, 267]; it appears that this also holds for Avalanche domains [91, 371].

Conversely, not preemptively seizing a malicious domain allows the botnet to respawn and continue its malicious operation: as the takedown does not remove the malware from infected machines, these will continue to establish contact with DGA domains. Once the botnet owners can obtain such a domain, the attackers can launch new attacks or spread malware to additional hosts. The takedown efforts, intended to permanently stop the malware, are then effectively spoiled.

Manually classifying all DGA domains is a resource- and time-consuming process, where due to 'decision fatigue' [136, 474], the mental effort in making repetitive decisions could lead to biases. Given the severe consequences of incorrect classifications, our goal is to develop an automated approach to the classification of DGA domains that performs with high accuracy, in order to relieve human investigators from manual effort as much as possible. At the same time, this does not preclude a manual review of those domains that are the hardest to classify or that could have the most significant effects. In the analysis of our approach in Section 6.5, we quantify how such a union of automated and manual classification can still lead to a significant reduction in required effort. Through such a reduction in manual effort and time, we can ensure the correctness of takedown decisions, thereby minimizing negative effects on website owners as well as end users.

6.3.2 Constraints for distinguishing malicious and benign domains

While our base goal is to distinguish malicious and benign domains, we cannot use previously proposed solutions as they rely on certain indicators that would not work for the Avalanche use case. Concretely, these indicators no longer hold for malicious domains (e.g. bulk registration), cannot be observed by us (e.g. detecting malware activity), or are counterproductive (e.g. alerting the attacker). Table 6.3 summarizes how the different contexts, goals and strategies of previous works do not fully satisfy our requirements.

Table 6.3: Overview of goals and strategies for the differentiation of benign and malware/DGA domains.

Context/Detection goal	Individual patterns	Proactive analysis	No active connections	Related work
Active malware domains within regular traffic	✗	✗	✓	[54, 55, 88]
Likely DGA domains within regular traffic	✗	✗	✓	[133, 432, 522]
Future malicious domains at registration	✗	✓	✓	[178, 221, 460]
Benign domains within known malware domains	✓	✗	✗	[262]
Benign domains within future DGA domains	✓	✓	✓	*Our work*

The reason is that the assumptions made in previous work no longer hold due to a different balance between malicious and benign domains: instead of detecting domains with clear malicious behavior among a (large) set of regular traffic, we assume that domains are malicious (they would be contacted by malware) and need to detect benign domains (i.e. accidental collisions). While in previous approaches, domains that do not exhibit strong indicators of maliciousness (offered by the former) are benign, the absence of such indicators in our use case means that we may not make such an assumption, and makes those previous approaches ineffective for Avalanche.

We translate these unique characteristics of the Avalanche takedown into three constraints. First, we need to take the characteristics of benign domains into account as well, by developing appropriate features that capture *individual differences in registration and configuration*. Second, as we cannot leverage ongoing malware activity itself, we need to develop features that allow for a *proactive analysis*. Third, attackers may not evade or detect data collection, so we may *not make any active connections* to domains in order to remain stealthy. In this section, we elaborate on these challenges and differences that make previous approaches ineffective for our use case.

Individual registration and configuration patterns Previous work often assumes that specific (bulk) patterns in the setup of domains indicates maliciousness.

For example, PREDATOR [221] relies on the observation that in order to evade blacklisting, malicious spam domains are registered in bulk (over 50% in groups of ten or more at one registrar in five minute intervals), causing these temporal clusters to be similar in infrastructure, lexical composition and life-cycle stage. In a similar spirit, Premadoma [460] relies on similarities in registrant data and the prevalence of malicious domains at specific facilitators (such as registrars) to detect sustained large-scale malicious campaigns. However, these patterns are no longer usable for our set of domains. Attackers only need to register one of the domains that the DGA outputs at a given time, so they no

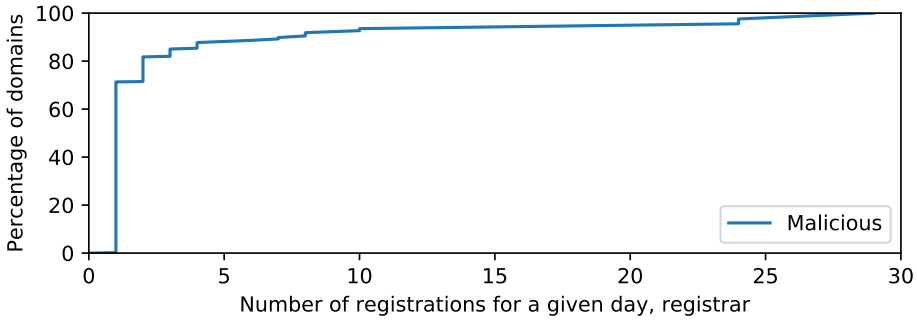


Figure 6.1: Cumulative distribution of registration counts for a given day and registrar, for malicious domains from the 2017 and 2018 iterations.

longer need to register domains in bulk, as is necessary for spam domains, also reducing the likelihood that they share e.g. registrars. Figure 6.1 confirms this: 93.5% of malicious domains in the 2017 and 2018 iterations of the Avalanche takedown are registered in clusters of fewer than 10 domains at their given registrar in one day (as opposed to the five minute interval in PREDATOR [221]). Moreover, the accidentally colliding benign sites do not have any relationship and will therefore not share any properties either.

Systems such as DeepDGA [522] and FANCI [432] detect DGA domains from linguistic patterns in their label. However, we know that all domains are either generated by a DGA or hard coded in malware, so it would be incorrect to use such patterns to categorize them as malicious.

In summary, because of the characteristics of our domain set (singular malicious and unrelated benign domains, all output by a DGA), many of the assumptions that the above approaches make on patterns that determine maliciousness are no longer valid. We must therefore resort to capturing more generic, common registration and configuration patterns for individual domains. These patterns should not only capture 'obvious' maliciousness, but also properties that indicate benignness.

Proactive analysis Previous work relies on observing ongoing malicious behavior: e.g. Exposure [88] leverages irregular DNS configurations and access patterns to detect 'domain flux' [230]; Pleiades [55] captures patterns in `NXDOMAIN` responses to DNS queries by active malware. These systems rely on ongoing malware activity that generates the analyzed traffic. Similarly, systems that use only the label to detect DGA candidates based on their appearance [133, 432, 522] need ongoing malware activity, otherwise infected hosts are not contacting malicious domains that are then visible in traffic.

Crucially, because malicious domains have to be taken down before they can cause any harm, we have to classify them proactively, i.e. before infected machines would actively query the malicious domain. This distinguishes our work from the above works, as we cannot analyze and rely on patterns within any (ongoing) malware activity. While we can

and do use features similar to those from previous systems, we are restricted to detecting patterns in registration, configuration, and regular traffic. Moreover, we already know that a DGA generated the domains that we have to classify, meaning that we start with an assumption that the domains are malicious.

No active connections to domains Internet measurements can be classified into two groups: passive collection, where already ongoing traffic is observed, and active collection, where new traffic is injected into the network. Notos [54] and Exposure [88] are examples of systems that analyze patterns in passively collected DNS queries. In contrast, Mentor [262] relies in part on website content features to measure positive domain reputation, requiring active and targeted data collection through crawling the domains.

While we have a similar goal to Mentor of detecting benign domains within presumably malicious domains, we avoid including features that require us to actively connect to domains. Malicious actors are namely known to detect active scanning and respond differently to appear more benign ('cloaking') [247], and could thus mislead our classification. More broadly, such probes could alert them of efforts to investigate and disrupt malicious infrastructures, allowing attackers to shift their approach or hide any traces to avoid repercussions [140]. A stealthier analysis without targeted active data collection therefore avoids endangering the effectiveness of ongoing investigations [88, 535].

6.3.3 Ground truth data

The advantage of our collaboration with law enforcement is that we can use their manual classification of benign and malicious domains from the takedown as a trustworthy source of ground truth. Previous studies mostly rely on publicly available blacklists and whitelists as the labeled ground truth [464], but malware blacklists have been found to contain benign parked or sinkholed domains and are ineffective at fully covering domains of several malware families [280], while lists of popular domains commonly used as whitelists can easily be manipulated by malware providers [291].

However, the real-world context of the Avalanche takedown affects the composition of our ground truth data. Concretely, our data set is relatively small, as seen in Table 6.2. Plohmann et al. [388] have seen a similarly small proportion of registered domains among DGA domains. We can expect this number to be small: malicious actors only need to register few domains, as the malware will try all DGA-generated domains; conversely, benign actors are less likely to be interested in using the often random-looking domains generated by the DGAs. Previous studies are able to evaluate their approach on much larger data sets, albeit self-constructed and arbitrarily selected. Nonetheless, training on a small data set is a challenge that prosecutors would also face, and our analysis is therefore valuable for informing them on the feasibility, constraints and benefits of an automated approach for such a practical use case.

6.3.4 Ethical considerations

We use the data set of the Avalanche takedown shared with us by our law enforcement partner. We augment this data with third-party data, avoiding unnecessary active probes of both benign and malicious domains. However, given the sensitivity of the former and commercial agreements for the latter, we cannot share this data with external parties. We release the data processing scripts and resulting models at `https://github.com/DistriNet/avalanche-ndss2020` to support reproducibility.

We assisted law enforcement agencies by applying our approach to the 2019 Avalanche iteration. While the use of machine learning for law enforcement purposes may be contested [385], human investigators may similarly make involuntary errors, e.g. due to 'decision fatigue' [136, 474].

6.4 Data set analysis and feature extraction

To determine a suitable takedown action for algorithmically generated domains (AGDs), we search for relevant features providing a full view of their properties over time. We then create a classifier that detects whether patterns in these properties are more likely to correspond to a benign or malicious domain without having to rely on ongoing malware activity.

In this section, we first analyze how different data sources can track different stages of the domain *life cycle* and we discuss the *insights* on how features capture contrasting properties of benign and malicious domains. Then, we select the final set of *features* and discuss the reasons for omitting certain features.

6.4.1 Life cycle of a domain

To correctly identify the intent of a domain registration, we need to observe patterns in the domain life cycle, as they indicate who obtained the domain, how they use it, and how they value it. For each identified step, we determine which relevant features capture the actions of the domain owner and list sources that track this information. Through our analysis, we can then ensure that our selection of features and data sets appropriately covers each step.

L1. Choice of the domain name The prospective owners of a domain (the registrants) must first choose the domain name that they want to purchase. Usually, the name is chosen to be easily memorized, sufficiently short, and representative of the service provided by the domain, but as malicious actors will need to produce domains in bulk, they will generate them automatically. The resulting names have a random or

patterned appearance that we can capture in lexical features on the label itself in order to automatically detect DGAs [429, 432, 522].

L2. Registration of the domain A registrant registers a domain through a registrar, typically paying a registration fee for at least 1 year [244] (although free and shorter offers exist [186] that tend to attract abuse [268]). The registrant identity, the registrar used, and the timestamps of the registration start and end are then made publicly available in the WHOIS database. We can then extract the registration patterns to distinguish benign and malicious sites [311]. Due to privacy concerns and regulations (e.g., the European General Data Protection Regulation), the publicly available identity of the registrant may be obfuscated: the real identity is then only available to the registrar and the top-level domain (TLD) registry. This data may be leveraged in collaborations with registries, e.g. for detecting malicious domains at registration time [460, 498].

L3. DNS configuration Once a domain has been registered, its entry in the Domain Name System (DNS) must be configured to allow discovery of its services using the domain name. The nameserver is passed onto the TLD registry and will appear in its zone files. The domain resource records configured in the nameserver zone file then become available for querying. Active DNS data sets (collected by e.g., OpenINTEL [489]) rely on scanning zone files or popular domains to obtain these records, while passive DNS data sets (collected by e.g., Farsight Security [174]) extract them from monitored DNS responses. Both types of data sets have been used to detect malicious domain registrations and activity [88, 270, 457].

L4. Setup of the service infrastructure The main purpose of a domain name is usually to provide a service for which an infrastructure needs to be set up. The records stored in DNS may reveal the hosting infrastructure or third-party service providers (e.g., cloud providers) from which actors that enable malicious activity can be derived [395, 534]. A scan of open ports accompanied by "banner grabs" may reveal provided services and the content available through the service may reveal its purpose. Such an operation requires active probing of the domain, which either can be executed ad hoc or is already performed regularly by e.g. Censys [150] and Project Sonar [398], whose scale enables analyses of botnet devices [56]. Furthermore, certificates obtained by the domain owner for their service may also be tracked in Certificate Transparency logs [286].

L5. Service activity Once the service is set up, end users can start interacting with it. Traffic to the service may be logged either at the server, the client, or in any network in-between. These logs can then be analyzed for multiple purposes. Malicious behavior can be detected and publicly shared in blacklists [280, 444, 534]. Commercial providers publish lists of the most popular websites that become base sets of seemingly benign domains [291]. The service may be crawled to populate search engine results or archive

web content [202]: the latter enables longitudinal analyses of malicious activity [46, 455, 534]. These methods can be combined to calculate risk scores for the domain [240].

L6. Service unavailability and domain expiration The unavailability of the services offered by the domain, either intentionally or unintentionally due to misconfigurations, may be detected by any of the previously discussed data sets depending on the type of disruption. Once a domain is no longer needed, it may expire: domains that are set to expire are often monitored for drop-catching [222], i.e., registering domains as rapidly after expiry as possible. Malicious actors also reuse previously expired domains to capitalize on the reputation of those domains [298, 524]. Alternatively, a service may be interrupted or a domain may be made unavailable for legal reasons, e.g., in takedown operations. As we study domains before they would be taken down, we do not consider this last step in our final feature set.

6.4.2 General insights

We want to design features that exhibit contrasting properties of benign and malicious domains and therefore provide a more accurate classification, while still acting within the constraints imposed by the Avalanche takedown use case (as outlined in Section 6.3.2). This requires insights into the generic differences in behavior of legitimate and malicious actors with respect to their domains. We choose our features to capture the following three characteristics:

i1. Likelihood of collisions Given that all domains are algorithmically generated, our target is to find "regular" (least random) looking domains as they are more likely to be a collision with a benign domain, which is opposite to other work that focuses on detecting DGAs solely based on how random their domain names appear [429, 432, 459, 522].

i2. Investment in the domain Obtaining and (validly) maintaining a domain requires an investment from its owner, both monetary for paying the registration fee and in effort for setting up DNS and WHOIS records correctly and installing services attached to the domain. While benign owners value their domains and are willing to make such an investment, the opposite is true for malicious actors: they want to set up a campaign with minimal cost and effort to maximize their revenue. Certain indicators imply high investment, such as long-term registration (benign domains tend to be older, while malicious domains tend to be registered shortly before the start of the validity period [88, 89, 189, 388]) or valid DNS and WHOIS records (invalid, obfuscated or repeated values hint at malicious practices [498]).

Table 6.4: Overview of the features used in our classifier. We indicate which outcome (benign or malicious) a higher or true value denotes and how the feature covers the domain life cycle and insights.

Set	#	Description	Type	Outcome	Life cycle step (Section 6.4.1)	Insight (Section 6.4.2)	Source
Lexical	1	Domain name length	Continuous	Malicious	L1. Domain choice	i1. Likelihood	[55]
	2	Digit ratio	Continuous	Malicious	L1. Domain choice	i1. Likelihood	[88]
Popularity	3	Number of pages found in Wayback Machine	Continuous	Benign	L5. Activity	i3. Popularity	*New*
	4	Time between first entry in Wayback Machine and takedown	Continuous	Benign	L5. Activity	i3. Popularity	*New*
	5	Time between first entry in Wayback Machine and start of malware validity period	Continuous	Benign	L5. Activity	i3. Popularity	*New*
	6-9	Presence in Alexa, Majestic, Quantcast, and Umbrella top websites rankings	Binary	Benign	L5. Activity	i3. Popularity	[302]
CT	10	TLS certificate found in Certificate Transparency logs	Binary	Benign	L4. Infrastructure	i2. Investment	*New*
WHOIS	11	Time between WHOIS creation date and start of AGD validity period	Continuous	Benign	L2. Registration	i2. Investment	*New*
	12	Time between WHOIS creation date and start of malware family activity	Continuous	Benign	L2. Registration	i2. Investment	*New*
	13	Time between WHOIS creation data and takedown date	Continuous	Benign	L2. Registration	i2. Investment	[189]
	14	Time between WHOIS creation date and WHOIS expiration date	Continuous	Benign	L2. Registration	i2. Investment	[262]
	15	Renewal of domain seen in WHOIS data	Binary	Benign	L2. Registration	i2. Investment	[221]
	16	Domain uses privacy/proxy service	Binary	Malicious	L2. Registration	i2. Investment	*New*
	17	WHOIS registrant email is a temporary/throwaway email service	Binary	Malicious	L2. Registration	i2. Investment	*New*
	18	WHOIS registrant phone number is valid	Binary	Benign	L2. Registration	i2. Investment	*New*
Passive DNS	19	Number of passive DNS queries	Continuous	Benign	L5. Activity	i3. Popularity	[302]
	20	Time between first and last seen passive DNS query	Continuous	Benign	L5. Activity	i3. Popularity	[302]
	21	Time between first seen passive DNS query and takedown	Continuous	Benign	L5. Activity	i3. Popularity	*New*
	22	Time between first seen passive DNS query and start of AGD validity period	Continuous	Benign	L5. Activity	i3. Popularity	*New*
	23-29	Presence of passive DNS query for resource record: `A`, `AAAA`, `CNAME`, `MX`, `NS`, `SOA`, `TXT`	Binary	Benign	L5. Activity	i3. Popularity	*New*
Active DNS	30	Time between first seen DNS record and takedown	Continuous	Benign	L3. DNS config.	i2. Investment	*New*
	31	Time between first seen DNS record and start of AGD validity period	Continuous	Benign	L3. DNS config.	i2. Investment	*New*
	32-36	Number of days DNS record was seen for resource records `A`, `AAAA`, `MX`, `NS`, `SOA`	Continuous	Benign	L3. DNS config.	i2. Investment	*New*

i3. Website popularity Establishing a website that attracts sufficient traffic and is therefore perceived as popular, requires significant effort in creating content and building an audience. Website popularity is therefore an indication of benignness: malicious actors will not make the effort of setting up real websites on dormant domains, especially as it is not required for the correct operation of botnets. Regular users as well as web crawlers are also unlikely to end up on these domains. Moreover, if the domain has not yet been generated by a DGA, its traffic is low or non-existent, so we can assume that any traffic that the domain draws is legitimate.

6.4.3 Summary of feature sets

We aim to capture the broadest view possible of the life cycle of the domains to classify, and select the features and the data sources that provide their values accordingly, further inspired by our general insights. While potentially useful, certain features are not applicable to our use case or would have unwanted consequences for required data collection or wider applicability of our approach. We elaborate on the reasons for not retaining these features in Section 6.4.4.

Table 6.4 gives a summary of the 36 features that we compute. We distinguish between six feature sets: for each feature set, we describe what it represents, which features it includes, how it is obtained, and how complete its coverage is. We indicate for each feature 1) whether it is binary or continuous, 2) whether our intuition is that higher or true values indicate a benign or malicious domain,[1] 3) which life cycle step from Section 6.4.1 it covers, and 4) which insight from Section 6.4.2 is illustrated.

For each domain, we know the start and end dates of their validity period, i.e. when their respective DGA would generate the domain. We also retrieve the date when a malware family started being active from DGArchive [388], where available.

Two *lexical* features capture the linguistic structure of the domain name. We compute the domain name length, as shorter domains tend to be more popular and expensive, and the ratio of digits in the domain name, as domains with more digits tend to be less readable. Both features discard the TLD.

Seven *popularity-based* features capture whether a domain hosts a website that appears to attract regular visitors. Three features use data obtained through the Wayback Machine API[2]: the number of unique pages captured on the domain, the time between the first capture of any page and the takedown, and the time between this first capture and the start of the AGD validity period.

Four features capture whether the domain is present at any point in time in the Alexa[3],

[1] Note that this is only an intuition—our classifier can detect edge cases that provide contrary evidence.
[2] `https://archive.org/help/wayback_api.php`
[3] `https://www.alexa.com/topsites`

Majestic[4], Quantcast[5], and Umbrella[6] top websites rankings. These rankings serve as an approximation of popularity from different vantage points: web browser visits, incoming links, tracking script/ISP data, and DNS traffic, respectively. Although they can contain malicious domains and are susceptible to malicious manipulation [291], we assume that presence in these lists still serves as a reasonable indication of benign intent. We retrieve historical data from an archive of historical top websites rankings [427].

One *Certificate Transparency* feature captures whether Certificate Transparency logs contain a certificate that was valid on the date of the takedown, i.e. whether the owner had obtained a TLS certificate for the domain. The feature in this set uses data obtained through an API from Entrust[7], which tracks Google Certificate Transparency logs [345]. Certificate Transparency logs have the most complete coverage of issued TLS certificates [490]. Recent browser policies that enforce logging further increase uptake [426].

Eight *WHOIS* features capture the registration cycle of a domain as well as registrant details. We base four features on the time between the WHOIS creation date and the start of the AGD validity period, the start of malware family activity, the takedown date, and the WHOIS expiration date respectively. For an additional feature, we compute whether the domain has been renewed at least once by the latest registrant, i.e. we find at least two records with different expiration dates.

We capture the validity of registrant data in three features. We determine if the domain uses a privacy/proxy service (replacing real registrant data with generic data) by checking for keywords (e.g. "privacy", "proxy") in the WHOIS registrant records. While legitimate users may prefer to use such a service to hide personal information [269], malicious domains also tend to use these services [122]. We also determine whether the WHOIS registrant email is a disposable address: as the email account can no longer be accessed after some time, this indicates that the owner does not consider the domain to be important. We test non-default/non-proxy email addresses against a manually curated list of disposable domains[8]. Finally, we check whether the WHOIS registrant phone number is valid: malicious actors would not want any trace leading to their real identity and therefore resort to fake (e.g., automatically generated) contact information. We test the validity of phone numbers using an API from numverify[9].

WHOIS-based features are based on historical data generously provided to us by DomainTools[10]. To observe long-term and renewed registrations, we obtain historical records spanning their full data collection period. The data reflects a state before the introduction of the European General Data Protection Regulation, so it contains more domains with publicly available contact details. We elaborate on the continued availability

[4] https://majestic.com/reports/majestic-million
[5] https://www.quantcast.com/top-sites/
[6] https://umbrella-static.s3-us-west-1.amazonaws.com/index.html
[7] https://www.entrust.com/ct-search/
[8] https://github.com/ivolo/disposable-email-domains
[9] https://numverify.com/
[10] https://whois.domaintools.com/

of such details in Section 6.6.2.

Eleven *passive DNS* features capture both the period and frequency of DNS resolutions for a particular domain, providing a viewpoint on both domain age and popularity. We retrieve the number of passive DNS queries: when more queries (for any resource record) have been made for the domain, the domain appears to be more popular. We base three features on the time between the first seen passive DNS query and the last seen query, the takedown date, and the start of the AGD validity period respectively. Finally, we record the presence of at least one passive DNS query for resource records `A`, `AAAA`, `CNAME`, `MX`, `NS`, `SOA`, and `TXT`: more (requested) record types with a value indicate proper domain setup and usage.

The features in this set use passive DNS data generously provided to us by Farsight Security[11]. We retrieve aggregated data spanning the full data collection period (i.e., since 2010 [174]). For each resource record value seen, the aggregated data contains the number of queries and the timestamps when it was first and last seen.

Seven active DNS features capture the availability of DNS records for a particular domain. We base two features on the time between the first seen DNS record and the takedown date, and the start of the AGD validity period respectively. We also record the number of days any DNS record value was seen for resource records `A`, `AAAA`, `MX`, `NS`, and `SOA`.

The features in this set use active DNS data generously provided to us by the OpenINTEL[12] project [489]. We cap the data period at 333 days (i.e. starting from January 1 of the relevant year). While OpenINTEL collects data actively, it complies with our requirement that we do not contact domains ourselves. Moreover, data collection is not targeted at specific domains, yet sufficiently comprehensive to also capture most of the registered Avalanche domains as it covers full zone files.

6.4.4 Omitted features

Given our use case of proactive takedowns, we cannot consider features that try to detect ongoing malicious operations directly, as the maliciously registered domain does not yet necessarily exhibit such behavior at the time of the takedown: malicious actors can leave these domains dormant right until a DGA generates the domain and infected hosts start contacting the domain. This means for example that we do not verify whether a C&C server is running on the domain and do not check malware blacklists.

Approaches for detecting AGDs, especially per single domain, are often based on lexical features that seek to discover patterns unlikely to occur in "human-generated" domain names [429, 432]. However, all of our candidate domains have been generated by a DGA, which leads us to use only a limited set of lexical features to find the domains that are more likely to be potential collisions (short and few digits).

[11] `https://www.farsightsecurity.com/solutions/dnsdb/`
[12] `https://www.openintel.nl/`

Detecting patterns from DNS logs [89] that indicate fast flux services [230], often used by command and control servers, is not applicable as the malicious domains would only start operating in fast flux during the validity period of the AGD.

Following our observation from Section 6.3.2 that bulk patterns do not apply for malware domains, we do not use approaches and features that rely on clustering domains [55] and batches of similar registrations [221], such as timing patterns or shared registrars.

The type of network could be an appropriate feature to take into account while the domain is active [89], with more trust in government or business networks hosting benign sites and domains in residential networks potentially being hosted by an infected machine. However, as a maliciously registered domain does not yet have to be actively malicious before the DGA generates the domain, its IP address can easily be set to a benign network (without the need for that network to actually host the domain) [327], thereby misleading our classifier.

Data collected through a crawl of candidate domains such as properties of the site content could indicate legitimately used domains [262]. However, following our stealth constraint from Section 6.3.2 and due to the need for historical data, we cannot do an active crawl of domains ourselves. We also cannot rely on existing third-party repositories of website crawls (e.g. the Internet Archive [510], Common Crawl [124] or Censys [150]): they do not provide historical data, do not crawl sufficiently regularly to capture recent data, do not have a consistent set of crawled domains and/or do not have sufficient domain coverage. Their data would therefore not be comprehensively representative of domain web content at the time of the takedown.

We do not include the malware family as a feature: as Avalanche provided domain registration as a service [140], we do not expect differences in behavior between the 21 supported malware families. Moreover, such a feature would go against our goal of capturing general differences in behavior between benign and malicious domains. We design the other features to represent distributions, for which the model can interpret the differences, whereas the malware family feature can only serve to refine the model for specific families. Finally, benign domains accidentally 'belong' to a certain malware family, so the feature is irrelevant in terms of registration behavior. We already capture relevant characteristics of the DGA in derived features such as the domain length that capture randomness in generated domains and therefore the likelihood of collisions.

We want to evaluate our approach as if it were deployed at the time of the takedown, so we do not use features for which we lack available historical data, as we would only be able to obtain the current state, which for malicious domains is post-takedown. They include the features that require active probing or data collection such as the website properties discussed earlier or the existence of search engine results for the domain, which could serve as an additional indicator of popularity. However, if they meet the applicable requirements and constraints, we can add such features in an actual takedown as we can then collect accurate data.

6.5 Analysis of machine learning-based classification

To evaluate to what extent machine-learning based approaches can reduce the effort of law enforcement to execute a takedown, we develop and evaluate a classifier that decides whether future DGA domains are likely to be benign or malicious. The goals of our analysis are threefold: we want to evaluate the raw performance of the classifier, but also gain insights into its decision-making process to finally thoroughly assess the benefits and limitations of automated approaches for domain classification. Moreover, given that not all data sources are equally easy to collect, we assess their impact on the correctness of our classification.

6.5.1 Experimental protocol

We first design an experimental protocol to determine the most appropriate machine learning-based solution and evaluate it in a way that is accurate and representative of real-world takedowns. Given the investigative setting and our intention to thoroughly analyze the resulting model, we restrict our selection of machine learning algorithms to those that are sufficiently interpretable. Moreover, as we systematically develop high-level features that capture the full domain life cycle, we do not require automated feature engineering. Therefore, we would not benefit from a deep learning approach and only face drawbacks from its increased complexity, so we do not consider it further.

Before classifying benign and malicious domains, we discard domains that were already sinkholed by security organizations to study botnet behavior. These organizations can sinkhole the domains either because they detect that botnet hosts are already contacting the domain (whose validity period therefore starts before and extends beyond the takedown date), or because they generate the domains output by the DGA upfront. The sinkholed domains can be considered neither a benign collision, as they do not host real content and may even mimic the malware C&C server, nor a registration made with malicious intent, as they will not communicate with actual malware. This means that they would confuse our model, and should be removed upfront by preprocessing the data. We detect sinkholed domains by matching DNS and WHOIS records with those of the sinkhole providers collected in SinkDB [14], by Alowaisheq et al. [46], and by Stampar et al. [461, 462]. Table 6.2 summarizes the distribution of domains across classes.

We execute our protocol with four machine learning algorithms: decision tree, gradient boosted tree, random forest, and support vector machine. We split data sets in a training and test set according to the considered iterations. When training and testing on the same iteration, we split the ground truth according to a 10-fold cross validation procedure. Otherwise, we construct the training and test sets from the separate iteration ground truths as applicable. We perform all model training and analysis using `scikit-learn` [380]. We elaborate on the different steps of this protocol in Appendix 6.A.

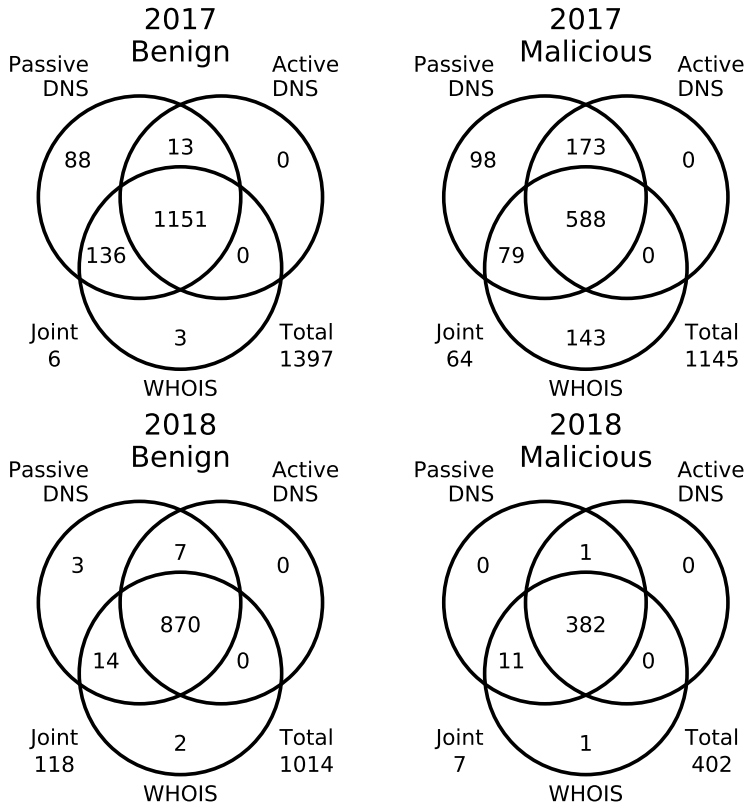


Figure 6.2: Number of domains where certain data sets are available, after removing sinkholed domains, for the 2017 and 2018 iterations. We separately mark the remainder of domains where only the joint data set (comprising lexical, popularity-based, and Certificate Transparency features) is available.

We run our experimental protocol for all domains of the 2017, 2018 and 2019 takedown iterations. We only evaluate performance with the manually labeled ground truth that we obtained from law enforcement for the 2017 and 2018 iterations (Section 6.3.3). In 2019, our model was used in the real-world classification effort, so a performance evaluation would be biased since we contributed to the ground truth.

As we want to measure the performance of our approach as if it were deployed at the time of the takedown operation, we use historical data that reflects the state of the domains as of each takedown, i.e. November 30 of each year. Data for the malicious domains collected after the takedown would refer to sinkholing and domain transfer infrastructure, making it a signal for maliciousness that would heavily bias our classifier.

As shown in Figure 6.2, we cannot obtain all data sets for all domains: this is because the third-party source could not collect relevant data (e.g. no WHOIS record is available or

Table 6.5: Performance metrics for the base ensemble model, varying the training and test set over the 2017 and 2018 iterations.

Training \ Test	Accuracy		F_1 score		Precision		Recall	
	2017	2018	2017	2018	2017	2018	2017	2018
2017	93.4%	84.3%	92.6%	73.4%	92.6%	70.8%	92.7%	76.1%
2018	76.1%	96.3%	70.9%	93.5%	78.6%	92.7%	64.6%	94.3%

the domain was never seen at passive DNS sensors). In order to still generate a prediction for all domains, we develop an *ensemble model*. We train a model for each combination of available feature sets, where a domain is included in the training set if at least those data sets are available. To classify a domain, we use the output of the model of the domain's available data sets.

6.5.2 Results

Given that we are the first to analyze the specific issue of preemptively deciding whether DGA domains are actually malicious or accidentally benign for a real-world takedown (which brings about certain constraints), we are not able to compare our performance results with previous work. Instead, we go beyond reporting basic metrics and critically examine how its performance translates into a real-world reduction in effort, whether our solution correctly captures differences between benign and malicious domains, and how much it depends on the availability of different data sets.

Model performance Appendix 6.B lists the relative performance of the four machine learning algorithms that we evaluate: we conclude that a gradient boosted tree classifier yields the best performance while still being sufficiently interpretable. We therefore analyze only its results.

We first train a *base* ensemble model, varying the training and test sets over the 2017 and 2018 iterations. From the performance metrics in Table 6.5, we can see that concept drift [519] occurs: performance drops when deploying our model across iterations instead of within. This suggests that over time, patterns that distinguish benign and malicious actors emerge or change, and these are therefore not captured by a model trained on only a single iteration.

We therefore develop an *extended* ensemble model, where we combine ground truth from a previous iteration with manual, *a priori* classifications of a subset of domains in the target iteration. This enables us to improve model performance by capturing the novel patterns in the new iteration, while still reducing manual effort overall.

We evaluate this extended model trained on all of the 2017 and part of the 2018 ground truth and tested on the remaining 2018 domains. Based on Figure 6.3, we empirically

Table 6.6: Performance metrics for models trained on the 2017 and (for the extended model) 15% of the 2018 iteration.

Ensemble model		Accuracy	F_1 score	Precision	Recall	FNR	FPR	Effort reduction
Base		84.3%	73.4%	70.8%	76.1%	23.9%	12.4%	100.0%
Extended	a priori	86.4%	78.6%	70.5%	88.6%	2.3%	2.0%	85.0%
Base	a posteriori	97.3%	95.3%	94.2%	96.5%	3.5%	2.4%	70.3%
Extended	a priori + a posteriori	97.6%	95.8%	94.3%	97.4%	2.6%	2.3%	66.2%

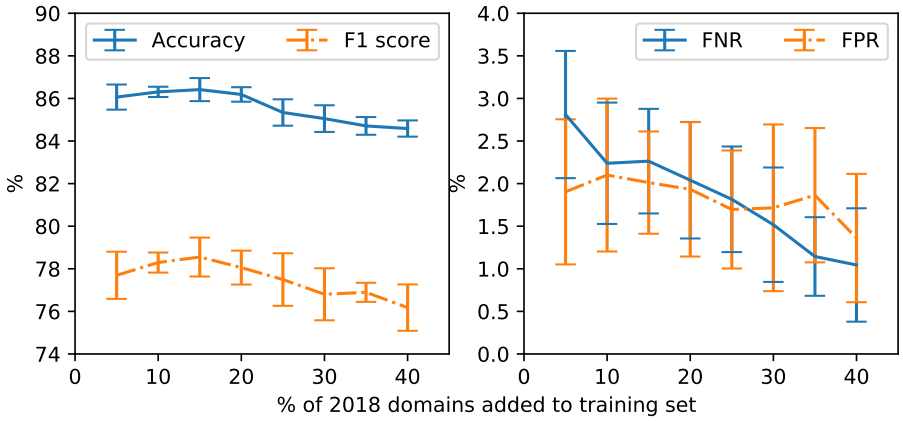


Figure 6.3: Performance metrics (mean and standard deviation) for the extended a priori ensemble model, trained on the 2017 and a varying part of the 2018 ground truth.

set the proportion of the 2018 ground truth that is (randomly) selected to be manually classified and added to the training set at 15%, as it represents the best trade-off between improved performance and limited additional effort. We repeat this random selection ten times and report average results. Table 6.6 shows that this extended a priori ensemble model improves on the base model.

However, some misclassifications still occur in this extended a priori model. The gradient boosted tree model outputs a score that reflects its confidence in its prediction. We can leverage these scores to develop a directed semi-automated approach: uncertain domains are manually investigated in more detail *a posteriori*. We examine how effective this approach is in further improving performance while still reducing investigative effort.

We explain this approach using the extended model for domains where all data sets are available, which allows us to simplify and visually support our explanation, but then apply it to the extended ensemble model. Figure 6.4 shows the false negative and positive rates as a function of the fraction of domains with a score below a certain value. By choosing a target maximum FNR and FPR, we can determine the lower and upper bounds on the

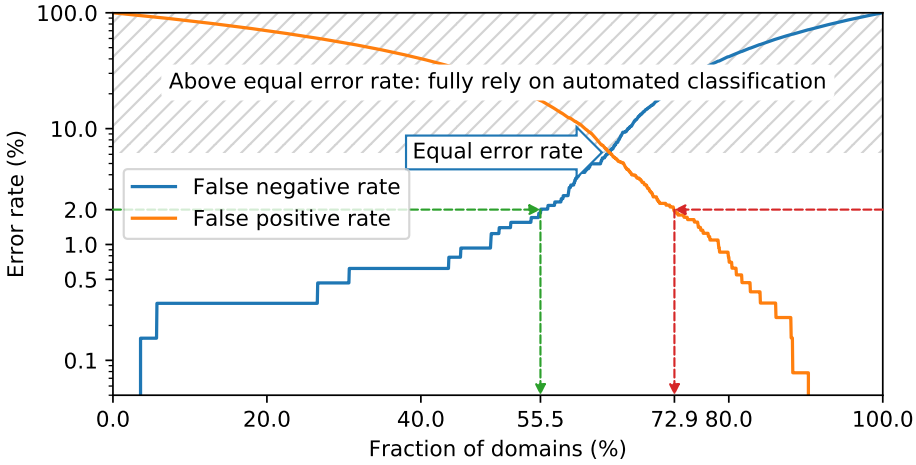


Figure 6.4: FNR and FPR as a function of the fraction of domains with a score below a certain value. By choosing the maximum error rate, we determine the fraction of domains that can be automatically classified.

maliciousness score; these bounds are determined based on the training set, so they do not necessarily reflect the exact actual error rates on the test set. Domains with scores within these bounds have to be verified manually, while domains with a lower and higher score are automatically classified as benign and malicious, respectively.

For the extended model on domains with all data sets available as represented in Figure 6.4, when setting a 2% error tolerance, 55.5% of domains have a maliciousness score below the lower bound set by 2% FPR (i.e. are benign), while $(100\% - 72.9\%) = 27.1\%$ of domains exceed the upper bound set by 2% FNR (i.e. are malicious). $55.5\% + 27.1\% = 82.6\%$ of domains therefore no longer need to be manually inspected. Only $72.9\% - 55.5\% = 17.4\%$ of domains still require further manual investigation.

When we apply this a posteriori approach to the extended ensemble model evaluated on all domains from the 2017 and part of the 2018 iteration (by choosing appropriate bounds for each component model), we obtain an accuracy of 97.6%; overall, the performance metrics in Table 6.6 indicate a very high performance. The effective FNR and FPR are 2.6% and 2.3%, comparable to the target error rate of 2%.

Overall, this approach reduces manual effort by 66.2%, accounting for the 15% of domains manually classified a priori. When the error tolerance is 1% and 0.5%, the fraction of automatically classified domains is 52.5% and 35.7% respectively. The score thresholds become very strict when very low error tolerances must be maintained, reducing the fraction of domains that can be automatically classified. The comparable effort reduction for an ensemble model trained on the 2017 and 2018 and tested on the 2019 iteration and a 2% error tolerance amounts to 76.9%, again achieving a significant reduction in manual effort.

Table 6.7: Importance scores of the top 10 features in the full feature set for the extended a priori ensemble model.

#	Set	Feature	Score
14	WHOIS	Time between WHOIS creation and expiration date	0.230
13	WHOIS	Time between WHOIS creation and takedown date	0.219
21	Passive DNS	Time between first passive DNS query and takedown	0.057
20	Passive DNS	Time between first and last seen passive DNS query	0.049
11	WHOIS	Time between WHOIS creation date and AGD validity	0.041
15	WHOIS	Renewal of domain seen in WHOIS data (Unknown)	0.040
34	Active DNS	Days DNS record was seen for resource record `MX`	0.040
15	WHOIS	Renewal of domain seen in WHOIS data (False)	0.037
31	Active DNS	Time between first seen DNS record and AGD validity	0.029
3	Popularity	Number of pages found in Wayback Machine	0.028

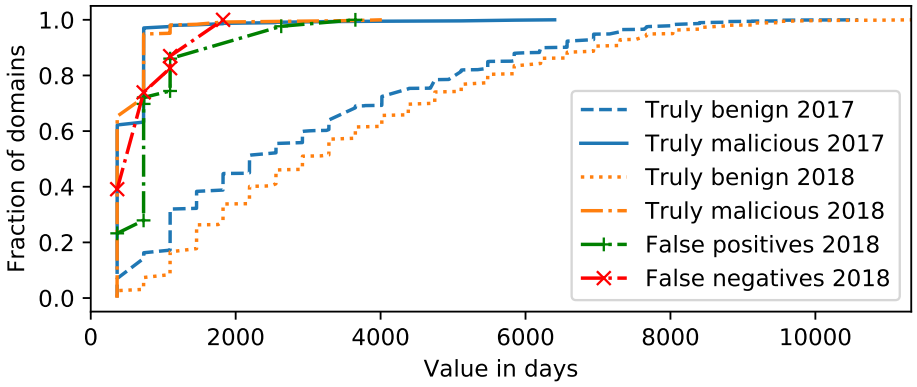
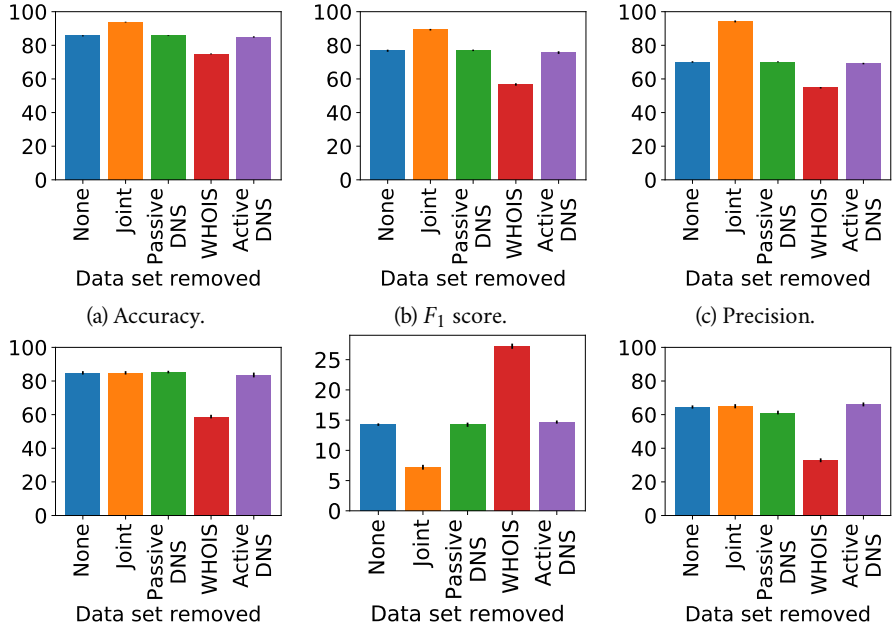


Figure 6.5: Cumulative distribution function of the values of benign, malicious, false positive, and false negative domains for the time between WHOIS creation and expiration date.

Feature analysis By using gradient boosted trees, we can measure how important individual features are to the overall performance. As we want to make an accurate assessment for the full feature set, we calculate importance scores for the extended model on domains where all data sets are available.

We show the ten most important features in Table 6.7 and find that they primarily capture the age and activity period of a domain. When malware creators want to evade our classifier, they would primarily want to influence these features. Figure 6.5 shows how the distributions of values for the most impactful feature (time between WHOIS creation and expiration date) are clearly distinct for benign and malicious domains. Misclassified benign domains (false positives) actually show a 'malicious' character, i.e. they are young; the malicious domains in our test set (from 2018) are never old, so other (but less expressive)



(a) Accuracy. (b) F_1 score. (c) Precision.

(d) Recall. (e) Equal error rate. (f) Reduction in effort (for 2% error rate).

Figure 6.6: Performance metrics (mean and standard deviation, in percent) of extended a priori ensemble models where one data set is omitted.

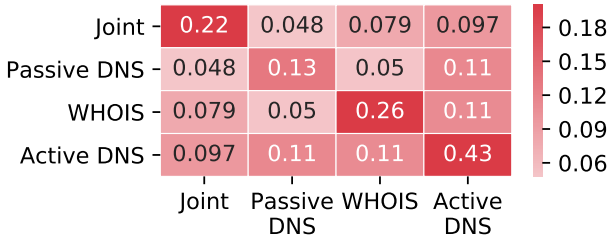
features impact whether they are classified correctly.

Consistent with our second insight from Section 6.4.2, time-based features are costly and difficult to evade: attackers have to register a domain name for a longer period of time, which translates into a higher monetary cost, and register it earlier, which is hard to achieve retroactively. In an extreme case, the domain name would have to be registered before the malware family becomes active.

Data set comparison We assess the impact of the availability of each data source on our performance starting from the extended a priori ensemble model, after which we retrain models with one feature set omitted each time. We join lexical, popularity-based, and Certificate Transparency features into a joint feature set, as they are the easiest to acquire and are always available, which leaves us with four feature sets: joint, WHOIS, passive DNS, and active DNS.

Figure 6.6 illustrates the performance of the models where one data set is discarded. We observe that missing WHOIS data has the most severe impact, significantly harming performance. Discarding the joint data set may actually improve performance, as its non-time-based features may lack sufficiently distinctive patterns, but it remains necessary

Table 6.8: Average covariance between features of one set, for the domains from the 2017 and 2018 iterations.



for domains that lack any other data set (but these are likely candidates for manual verification).

Missing passive or active DNS data has a less pronounced effect. We find some degree of redundancy between passive and active DNS data, as their time-based features in particular represent similar concepts and are therefore intuitively dependent. We confirm this effect with the covariance between feature sets shown in Table 6.8: passive and active DNS data are relatively highly correlated with each other.

This effect means that passive and active DNS (as well as WHOIS) data all capture important and hard-to-evade time-based patterns, but that one missing data set can be substituted by the others without a significant loss in performance. This becomes important when considering that data sets such as WHOIS that lead to better performance may come with a significant cost to acquire. In Section 6.6.2, we elaborate on the implications of our findings on future takedown operations.

Conclusion We find that an approach combining primarily automated classification and targeted manual investigation across multiple iterations achieves the best compromise of high accuracy and low manual effort, with less than 3% mistakes. This reduces investigative effort by up to 76.9%, depending on the tolerated error rate, freeing up time to focus on those domains that are the hardest to classify.

Our analysis of features and data sets shows that time-based features are the most important ones, which at the same time increases the cost and difficulty of evading our classifier. However, our performance depends on data sources with a high cost of acquisition, in particular WHOIS data. We continue our discussion of these aspects in the next section.

6.6 Discussion

In this section, we elaborate on the factors that may influence the applicability of our approach to future takedowns. We first explain how a high cost and effort for attackers

complicates the evasion of our classifier and may therefore discourage malicious actors. We then highlight how recent developments in the availability of data sets may have a negative impact on the performance of our approach.

6.6.1 Evasion

Previous work [221, 311] pointed out that attackers may develop bypasses to mislead a classifier like ours and therefore evade detection and subsequent takedown of their malicious domains, especially as we cannot rely on detecting the malicious activity that would be required for the correct functioning of the botnet. We discuss potential evasion strategies and how difficult they are for malicious actors to deploy. This proactive analysis allows for anticipating changes in attacker behavior, developing additional features that are even harder to circumvent and implementing infrastructural measures that complicate evasion.

Features that leverage the properties of the DGA itself, such as lexical features, can be evaded by redesigning DGAs. While it is feasible to carefully engineer DGAs to be more resilient against detection [459], such a DGA should generate domains that appear very similar to benign domains (e.g., only short domains). This yields a higher risk of collisions and fewer domains available for registration, endangering uninterrupted control of the botnet.

Popularity-based features require setting up a website for discovery by web crawlers, and generating traffic, or at least the appearance thereof. Website popularity rankings can easily be manipulated at scale [291], allowing attackers to insert their domains and appear as benign. If malicious actors can have a presence within the networks where passive DNS data is collected, they could also insert DNS traffic that makes the domain appear regularly visited. Given that the attackers control their infected machines, the botnet itself could be leveraged for this purpose. However, as the traffic of infected machines can be monitored, these queries can be detected, revealing those domains that the malicious actors have registered upfront. Finally, the presence of certain DNS resource records can be forged by inserting fake records, but as some records require values of a specific format, their validity could be verified, as maintaining valid records requires more effort.

Given recent efforts to increase the ubiquity of TLS encryption by making free and automated TLS certificates available [26], malicious actors can relatively easily obtain them for malicious domains and therefore appear in Certificate Transparency logs. However, such a process still requires additional effort that is not strictly necessary for the correct operation of the C&C server. While the choice to obtain a paid certificate indicates a willingness to invest in the domain (and therefore suggests benignness), the use of a free certificate does not necessarily imply maliciousness.

Features that consider the age of a domain can be thwarted by registering malicious domains (long) before they become valid. However, it requires prolonged registrations and the corresponding payment of registration fees, which runs counter to minimizing

the cost of the malicious campaign. Moreover, the longer a domain with malicious intent has been registered, whether active or dormant, the more susceptible it is to being blacklisted/taken down or to the attackers being identified.

Acquiring and managing domains may incur a significant (manual) effort. If the process is automated, certain registration patterns can emerge that make it easier to identify the maliciously registered domains [460, 498]. Malicious actors might attempt to compromise existing or reuse expired domains to exploit the (residual) trust in these domains [298] (for example their age). However, it would require even more effort, as they would need to find eligible domains, attempt to compromise them or monitor their expiration status to take them over at the right time, and finally deploy the malicious operation. As domains are randomly generated by a DGA and often have a short validity, the likelihood of success is low.

To circumvent features that use WHOIS registrant records, malicious actors could insert forged yet realistically-looking data. However, if these records are automatically generated, detection becomes feasible and accurate [460, 498]. Manual effort in creating fake records quickly becomes infeasible given the need to keep registering domains as they become (in)valid.

In summary, while the publication of features allows for an attacker to develop techniques to evade them, many of these would go against the goal of malware operators to set up these domains with low effort and at low cost. Moreover, if the attacker behavior would significantly shift, other evasion countermeasures and detection strategies remain available, although they might require increased effort and involvement by relevant stakeholders. Finally, we find time-based features to be the most important ones: they are particularly costly and hard to evade.

6.6.2 Availability of data sets

Our features come from different data sources that each present their own issues in terms of acquisition, affecting not only law enforcement but also adversaries seeking to evade the model. Moreover, our evaluation of the importance of different data sources for correctly classifying domains shows that the data sets that contribute the most to our model's performance have a significant cost in terms of money and effort.

WHOIS data in particular provides the highest accuracy, but obtaining it may be challenging. From a technical standpoint, WHOIS data is not machine-readable nor has a standard format [135], so it requires (sometimes manual) parsing. Moreover, access is rate limited [305].

Public availability of WHOIS data is also affected by privacy concerns [408] as well as strict limitations on the collection and dissemination of personal data due to privacy regulations. This triggered ICANN to adopt the "Temporary Specification for gTLD Registration Data", which allows generic TLD registries to redact personal data in WHOIS records,

while having the intent to provide vetted partners such as law enforcement agencies with privileged access [245]. As a result of the European General Data Protection Regulation, European country-code TLD registries have also started to withhold personal data [142]. Security researchers have voiced concerns that the unavailability of such data to them could significantly hamper efforts to identify and track malicious actors [180, 387].

Passive DNS data collection may also have privacy implications [270], and requires sufficient storage and processing resources. Active DNS data collection has similar storage and resource needs, especially to ensure that records are updated sufficiently frequently. The coverage of both data sets also depends on cooperation of third parties: passive DNS requires access to recursive resolvers ideally deployed all over the world, and active DNS collection often relies on zone files that must then be shared by registries. Although law enforcement may gain more extensive access, they may be more limited in terms of resources, and delays in procedures to obtain data may hamper swift action. Conversely, commercial providers that can deploy more extensive resources may not be able to access more sensitive information. Finally, from a cost perspective, these commercial providers may charge significant amounts, especially for historical data.

We see that our approach becomes less effective if certain data sets would be unavailable, and our discussion shows that comprehensive coverage of data sets comes at great cost. However, we can still achieve reasonable performance even with missing data, and we see that data sets are partially correlated. The continued availability of these data sets is therefore important to counter future malicious operations, but not to such an extent that their absence would be disrupting the effectiveness of takedowns.

6.7 Related work

Classifiers for detecting malicious domains Numerous works have addressed the problem of designing classifiers to distinguish benign from malicious web pages and domains. Ma et al. [311] classified malicious URLs based on lexical and host-based features, comparing multiple feature sets and classifiers. Felegyhazi et al. [178] designed a classifier seeded with known malicious domains that uses DNS and WHOIS data. Antonakakis et al. [54] proposed Notos, which outputs a reputation score based on the determination of the reputation of domain clusters obtained from network properties, DNS data, and the ground truth on benign and malicious domains. Bilge et al. [88, 89] proposed Exposure, which uses DNS-based and domain name features to detect domains contacted by infected machines within passive DNS traffic. Frosch et al. [189] proposed Predentifier, which combines passive DNS, WHOIS, and geolocation data to detect botnet command and control servers. Hao et al. [221] proposed PREDATOR, a classifier for malicious domains based on features available at the time of registration and the identification of batch registrations. Spooren et al. [460] developed Premadoma, a model to detect malicious domains at the time of registration, leveraging features based on infrastructural reputation and registrant similarity, and discussed the challenges and tactics for deploying the model in an operational setting. Machlica et al. [312] created a

model that uses two levels of classifiers to improve detecting malicious domains using lexical and traffic-based features. Kidmose et al. [263] and Zhauniarovich et al. [535] surveyed approaches to detecting malicious domains from (enriched) DNS data.

Classifiers for detecting algorithmically generated domains Earlier work in detecting algorithmically generated domains (AGDs) identified clusters of likely candidates. Yadav et al. [526, 527] evaluated several statistical measures for classifying groups of domains as algorithmically generated or not based on character distributions within the domain names and the IP addresses to which they resolve. Yadav and Reddy [525] applied similar statistical measures on successful and failed domain resolutions. Antonakakis et al. [55] proposed Pleiades, which clusters non-existent domains based on character distributions within the domain names and on the querying hosts, using the strategy on DNS traffic from large ISPs to discover six DGAs that were unknown at that time. Krishnan et al. [276] detected hosts in a botnet by analyzing patterns in DNS queries for non-existent AGDs through sequential hypothesis testing. Mowbray et al. [348] detected hosts that query domains with an unusual length distribution, deriving 19 DGAs of which nine were previously unknown.

Later work moved towards detecting AGDs per single domain name. Schiavone et al. [429] proposed Phoenix, which uses linguistic features to detect potential AGDs, afterwards using linguistic, IP-based and DNS-based features to cluster domains and extract properties of the DGAs that generated them. Abbink and Doerr [8] and Pereira et al. [384] highlighted how most classifiers focus on detecting the randomness in AGDs and are therefore not able to correctly classify dictionary-based DGAs, and proposed new methods for detecting such DGAs. Multiple deep learning-based approaches have since been proposed [447]. Spooren et al. [459] found one such deep learning model by Woodbridge et al. [522] to outperform the human-engineered features of the model by Schüppen et al. [432].

Takedowns of botnet infrastructures Previous coordinated takedowns of botnet infrastructures have been studied to evaluate their effectiveness over time in preventing further abuse. Nadji et al. [353] presented rza, a tool that uses a passive DNS database to analyze and improve the effectiveness of botnet takedowns. They evaluated the tool for three malware families and found mixed long-term impact of takedown operations. Asghari et al. [62] analyzed the institutional factors that influenced the cleanup effort of the Conficker worm, finding that cleanup was slow and that large-scale national initiatives did not have a visible impact. Shirazi [437] surveyed and taxonomized 19 botnet takedown initiatives from 2008 to 2014. Plohmann et al. [388] analyzed the structure of DGAs for 43 malware families and variants, and analyzed registrations of their AGDs, finding domains missed in takedowns, families for which few domains were sinkholed, and slowness in seizing AGDs registered by malicious actors. Alowaisheq et al. [46] studied the life cycle of takedown operations across sinkholes and registrars based on passive DNS and WHOIS data, finding several flaws that would allow malicious actors to regain control of some sinkholed domains. Hutchings et al. [239] provided insights into the effectiveness

of takedown efforts by interviewing key actors, finding that law enforcement faces more challenges than commercial enterprises in effectively carrying out takedown operations.

6.8 Conclusion

Taking down the domains that compromised machines use to communicate with command and control servers is an effective measure to disrupt botnets such as Avalanche. However, law enforcement must take care not to affect any legitimate domains that happen to collide with algorithmically generated domains. For Avalanche, prosecutors manually conducted this classification process, requiring large amounts of time and effort as well as allowing for human error.

We therefore develop an automated approach for classifying benign and malicious registered DGA domains, within the constraints of the real-world takedown context that make previous approaches inapplicable: we cannot rely on bulk patterns, detecting ongoing malware activity or actively connecting to domains. We propose a hybrid model that balances automation with manual classification to achieve a higher performance as well as vastly reduce investigator effort. We develop and evaluate our approach to represent the Avalanche takedown most truthfully, such that our results and findings reflect the utility of automated domain classifiers in a real-world takedown scenario, such as for our contribution to the 2019 iteration.

Given the increasing number and size of cybercrime operations, automated tools can assist law enforcement investigators in avoiding any harmful impact of their operation, especially on uninvolved legitimate parties. These tools will allow them to stay one step ahead of malicious actors and impair their activities with the goal of shielding end users from any harm.

6.A Machine learning protocol

Machine learning algorithms are trained on a training set Tr and evaluated on a test set Te. As explained in Section 6.5, if we need to train and test on the same iteration, we split using a k-fold cross validation procedure: the data is split in k folds, with every fold being used once as the test set, while we use the $k-1$ others for training, and finally, we average results over k experiments. We set k to 10. The advantage of using cross validation is that we can reduce bias in the composition of the selected training and test set, even with a relatively small data set.

Most ML algorithms have different hyperparameters to tune. Tuning on the test set would lead to highly biased results. Therefore, we have to split the training set Tr into a set for training Tr' and another one for validation V. We again use a 10-fold cross

validation procedure. We treat and calculate the upper and lower bounds for the extended a posteriori model as hyperparameters.

We evaluate the following performance metrics over the test set:

$$accuracy = \frac{tp + tn}{tp + tn + fp + fn} \tag{6.1}$$

$$precision = \frac{tp}{tp + fp} \tag{6.2}$$

$$recall = \frac{tp}{tp + fn} \tag{6.3}$$

$$F_1 = 2 * \frac{precision * recall}{precision + recall} \tag{6.4}$$

where tp, tn, fp, fn stand for the number of true positives, true negatives, false positives and false negatives, respectively. Malicious domains are considered positive, benign domains are negative. Precision represents the fraction of samples identified as malicious that are actually malicious, while recall represents the fraction of malicious samples that were correctly identified. The F_1 score summarizes these two metrics, and is a superior metric compared to accuracy when dealing with unbalanced datasets, therefore we optimize for it.

Due to incompleteness of our data sets (e.g., WHOIS records not containing a parseable phone number), certain domains have missing feature values. We impute them (i.e., substituted them with plausible values to avoid bias) as follows (the feature numbers correspond to those defined in Section 6.4.3):

- No Wayback Machine data: feature values (3-5) are set to zero as no data means that the Wayback Machine has not found any page on the domain, suggesting unpopularity.
- No WHOIS timestamps: feature values (11-14) are set to the mean, as no data implies that data could not be parsed or retrieved, not that the data does not exist (e.g., all domains have a registration date). By using the mean, we do not attach any statistical meaning to the absence of data and do not skew the distribution.
- Less than two WHOIS records: the renewal feature (15) gets a third value that indicates that only one historical WHOIS record was available (preventing a comparison of expiration dates).
- No WHOIS registrant records: features that rely on an address, an email address, or a phone number (16-18) get a third value that indicates that we do not have a value for the corresponding field.
- No passive or active DNS data: continuous feature values (19-22, 30-36) are set to zero and binary feature values (23-29) to false as no data means that DNS records for the domain were never queried, suggesting unpopularity.

6.B Evaluation of machine learning algorithms

Table 6.9 presents the performance metrics of the machine learning algorithms that we evaluate in Section 6.5.2, for a base ensemble model trained and tested on the initial 2017 iteration. The results show that gradient boosted trees consistently outperform the other ML algorithms.

Table 6.9: Performance metrics of the evaluated machine learning algorithms.

Metric	Decision Tree	Gradient Boosted Tree	Random Forest	Support Vector Machine
Accuracy	88.6%	93.4%	92.8%	86.4%
Recall	86.6%	92.7%	92.6%	77.9%
Precision	87.8%	92.6%	91.5%	90.6%
F_1 score	87.2%	92.6%	92.0%	83.8%

7

An Audit of Facebook's Political Ad Policy Enforcement

This chapter is based on the homonymous paper published in the proceedings of the 31st USENIX Security Symposium (2022) [287]. This work was co-authored with Laura Edelson, Tom Van Goethem, Wouter Joosen, Damon McCoy, and Tobias Lauinger.

Major technology companies strive to protect the integrity of political advertising on their platforms by implementing and enforcing self-regulatory policies that impose transparency requirements on political ads. In this paper, we quantify whether Facebook's current enforcement correctly identifies political ads and ensures compliance by advertisers. In a comprehensive, large-scale analysis of 4.2 million political and 29.6 million non-political ads from 215,030 advertisers, we identify ads correctly detected as political (*true positives*), ads incorrectly detected (*false positives*), and ads missed by detection (*false negatives*). Facebook's current enforcement appears imprecise: 61% more ads are missed than are detected worldwide, and 55% of U.S. detected ads are in fact non-political. Detection performance is uneven across countries, with some having up to 53 times higher false negative rates among clearly political pages than in the U.S. Moreover, enforcement appears inadequate for preventing systematic violations of political advertising policies: for example, advertisers were able to continue running political ads without disclosing them while they were temporarily prohibited in the U.S. We attribute these flaws to five gaps in Facebook's current enforcement and transparency implementation, and close with recommendations to improve the security of the online political ad ecosystem.

Table 7.1: Summary of (in)correctly classified ad counts, across undeclared political ads as labeled by Facebook or by us, within a 14-day observation period after an ad's first activity.

Detected as political by Facebook		Not detected by Facebook
40,191 *	**32,487** *	**116,963** §
False positive (Section 7.5.2)	True positive (Section 7.5.1)	False negative (Section 7.6.3)
Not political	Actually political	

Precision: 0.45 *Recall*: 0.22 F_1 *score*: 0.29

* Across all advertisers worldwide; estimate based on 55% FP rate in U.S.
§ Across political advertisers worldwide.

7.1 Introduction

Online political advertising is a powerful tool for enabling engagement in the political process, but with this power comes the risk of abuse that can harm the integrity of the democratic process. Scrutiny of major online advertising platforms intensified due to foreign interference in the 2016 U.S. elections [401] as well as broader concerns on disinformation, voter suppression, and inauthentic behavior [409]. As government regulation has failed to adapt [252, 295], oversight on online political advertising has fallen largely to the platforms themselves [170, 252]. Platforms therefore developed self-regulatory policies [295] that include verifying and revealing advertisers' identity [197, 233], creating public archives of political ads [11], or even banning political ads altogether [3, 252].

A baseline requirement for platforms to protect integrity and reduce harm is then to properly identify advertisements that seek to influence public opinion, and adequately enforce their policies on those ads and their advertisers. Failing to do so correctly, rapidly, and consistently leaves an opportunity for ill-spirited advertisers to impede public scrutiny, spread violating content, and evade restrictions on political ads. Conversely, well-meaning advertisers are disadvantaged if their ads are unduly made unavailable due to incorrect enforcement, or if they (over-)comply with policies while others do not [297], especially when policies are unclear or ambiguous. Given the large number of submitted ads, platforms usually deploy automated methods for policy review, complemented by human review when needed [109, 161, 321].

In this paper, we audit whether Facebook makes accurate enforcement decisions for ads that may be in scope of its political ad[1] policies, but were not declared as such by the respective advertisers. Facebook is the most popular social media platform worldwide among users [259] and advertisers [463], and its political ad policies and transparency

[1] In this paper, we use 'political ads' as shorthand for ads in scope of Facebook's policy, i.e., "ads about social issues, elections or politics" [23].

are considered to be among the most developed for major technology companies [295, 451], allowing us to analyze the effectiveness of self-regulation through one of the most advanced deployments. We build a novel large-scale data collection pipeline that retrieves all currently active ads running on Facebook's core advertising platforms[2] from the *Ad Library*, its ad transparency tool. Our comprehensive and representative data set contains 4.2 million political and 29.6 million non-political ads from all 215,030 pages[3] that ran political ads during the second half of 2020 and beginning of 2021, covering major elections in the U.S. and Brazil. We analyze the prevalence of ads that Facebook correctly detects to violate policies after they start running (*true positives*), ads that Facebook detects but are not political (*false positives*), and ads that Facebook fails to detect even though they are political (*false negatives*).

In prior work, the Ad Library has been used to study advertisers evading Facebook's transparency requirements [157], while other research sought to quantify enforcement errors through anecdotal evidence [153, 440, 484, 509], or through crowdsourced [439, 456] or self-published [321] ads; however, these studies inherently cover only a small sample of ads. To the best of our knowledge, no previous study has quantified the *performance* of Facebook's political ad policy *enforcement* in detecting non-compliance *at a large and representative scale*. A study such as ours is essential to understanding whether Facebook's current self-regulation effort is sufficient to maintain the integrity of its political ad ecosystem.

Overall, we find that policy violations detected after an ad starts running represent a small share (1.7%) of political ads on Facebook. Detection happens rather quickly; yet it is worth noting that these violating ads failed to be detected during Facebook's initial ad review, which allowed them to accumulate over 2 billion user impressions before being taken down. Unfortunately, this detection of violating ads seems to have little visible impact on advertisers. Despite a history of violations, we observe that the top violating advertisers were able to continue running new ads and accumulate more violations for long periods of time, even while political ads were banned in the U.S. [3, 284, 435].

Ambiguities in Facebook's policies and flaws in Facebook's existing detection appear to cause many unrelated ads to be incorrectly labeled as political: We estimate that among U.S. advertisers, 55% of ads detected as "political" by Facebook are in fact false positives. Conversely, we identify 39% of advertisers in Facebook's Ad Library Reports as clearly political. While such advertisers are subject to a blanket rule in Facebook's ad policy requiring them to declare *all* their ads as political, these pages ran a total of 116,963 ads that were not declared as political and not detected by Facebook. In addition, significant differences in the rates of undetected ads arise between countries: While performance is best in the United States at 0.85%, Facebook may fail to detect up to 45% of undeclared political ads in other countries.

When considering only the running ads where Facebook needed to make an enforcement

[2]Facebook, Instagram, Messenger, and the Audience Network.

[3]An *advertiser* runs ads from their Facebook *page* [11]. In this paper, we use 'advertiser' and 'page' interchangeably.

decision, that is, ads not voluntarily disclosed by their advertisers, we find 61% more ads that are missed than are detected by Facebook within 14 days, and 55% of detected ads are likely false positives (Table 7.1). With more errors than correct decisions, Facebook's current enforcement approach appears inadequate: users are left vulnerable to ads that seek to influence their opinion without proper disclosure, while legitimate advertisers regularly see their ads unjustly taken down. We attribute these flaws to insufficient attention for an advertiser's political intent, the possibility for advertisers to continue running violating ads, inadequate localization in many countries, and ambiguity in policies, worsening transparency. Based on these observations, we make a number of recommendations to improve policy enforcement (Section 7.7).

In summary, our main contributions are:

- We develop a novel data collection pipeline through which we obtain a comprehensive and representative view on all active political and non-political ads running between July 2020 and February 2021 across 215,030 pages (Section 7.4).
- From an ad-level perspective, we find 1.7% of all "political" ads to have been detected post-hoc by Facebook, but detection is imprecise: we estimate through manual labeling that in the U.S., 55% of detected ads were incorrectly marked as political (*false positives)* and taken down (Section 7.5).
- From an advertiser perspective, we find that detection of violating ads does not appear to prevent future violations, and that Facebook misses 116,963 ads from clearly political advertisers (*false negatives)*, with considerably worse performance outside the United States (Section 7.6).
- We identify five factors where our findings suggest that Facebook's current enforcement and transparency implementation is lacking, and suggest improvements that would strengthen enforcement and improve the security of the online political ad ecosystem (Section 7.7).

7.2 Background

7.2.1 Political ad policy

Facebook imposes increased authenticity and transparency requirements for "ads about social issues, elections or politics," by requiring advertisers to confirm their identity and location and declare who funded the ads. These requirements are only mandatory and therefore proactively or reactively enforced in around[4] 60 countries and territories for ads about social issues, elections or politics, and in around 60 additional countries for ads about elections or politics only [64], with these sets of countries expanding over time. In all other countries, advertisers are currently "strongly encouraged" to get authorized and declare ads, but this is voluntary and not enforced [64].

[4]Lists of countries are inconsistent between the web portal [11] and documentation [64].

Facebook considers ads to be "about social issues, elections or politics" if they are [24]:

- "made by, on behalf of or about a candidate for public office, a political figure, a political party, a political action committee or advocates for the outcome of an election to public office; or
- about any election, referendum or ballot initiative, including "get out the vote" or election information campaigns; or
- about any social issue in any place where the ad is being run; or
- regulated as political advertising."

Facebook further specifies 'social issues' as "sensitive topics that are heavily debated, may influence the outcome of an election or result in/relate to existing or proposed legislation" and requires disclosure for these "social issue ads that seek to influence public opinion" [10]. Facebook defines a list of top-level 'social issues' per country (where applicable), which can change over time [10]; Facebook further clarifies these topics with examples of ads that are in and out of scope [232].

Before an advertiser can run ads about social issues, elections or politics in an applicable country, they must complete the *authorization process* there and confirm their identity and location [24, 114, 161, 197]. Once authorized, they can create 'disclaimers' to indicate which funding entity (individual, page or organization) paid for a given ad [92, 129, 161]. When running a political ad, the advertiser must then select it as a "Special Ad Category" [114] and add a disclaimer [114, 161].[5]

As shorthand, whenever we mention 'political ads' we refer to "ads about social issues, elections or politics" that were properly *declared* (i.e., having a disclaimer) or *detected* (i.e., lacking a disclaimer but marked as political by Facebook).

7.2.2 Policy enforcement

Facebook requires advertisers to self-determine that an ad is in scope of its ad policy on social issues, elections or politics, but also reviews any other submitted ad for policy compliance [161]. This "relies primarily on automated review (artificial intelligence) [...] and, in some cases, [they] have trained global teams to review specific ads" [161]. If an undeclared ad gets caught during this initial review, it never runs and is not archived in the Ad Library; the attempted violation will never be publicly known. This paper excludes such early detections.

If an undeclared ad passes review and is running, it can still "be flagged by AI or reported by [Facebook's] community" as political [161]. Facebook then "disapproves" the ad retroactively, meaning they deactivate the ad, so it is no longer shown to any user. This is the type of ad detection that we study in this paper. Facebook also archives the violating ad in the Ad Library with a message that "this ad ran without a disclaimer," regardless of whether the advertiser completed the ad authorization process [9, 11]. While the ad will

[5]Certain Facebook-vetted news publishers are exempt from declaring ads even if their content is political but not opinionated [16, 24].

remain publicly archived even when inactive, it will therefore never be known who paid for the ad. Violating pages may also be restricted from running new (political) ads or be disabled [24].

7.2.3 Transparency tools

Facebook emphasizes transparency as a means to hold them and their advertisers accountable [161], enabling users to be aware of who is trying to influence them as well as enabling journalists, organizations, and researchers (including us) to audit online political advertising [161]. To support this transparency, Facebook provides three core tools [515]:

1. The *Ad Library* [17] is a web portal where users can search all currently active ads for any Facebook page in any country, as well as all active and inactive ads about social issues, elections or politics. Only for the latter, provided metadata includes the disclaimer provided (if any), the identity of an authorized advertiser and how this was verified, and binned estimates of ad spend, reach, and impressions. A non-political ad disappears from the Ad Library once it becomes inactive; a political ad is archived for 7 years [11]. Section 7.A shows how the web portal displays ads.
2. The *Ad Library API* [18] provides an interface for automated queries for all active and inactive ads about social issues, elections or politics for any page in a given country.
3. The *Ad Library Report* [19] aggregates advertiser data for all ads about social issues, elections or politics for countries where Facebook requires disclosure, listing all pages with at least one political ad in the chosen time span.

7.2.4 Related work

Prior work used crowdsourced or self-published ads to analyze the correctness of Facebook's political ad policy enforcement. Silva et al. [439] developed a system to crowdsource Facebook ads in Brazil and classify them as political using a supervised machine learning model. Across 38,110 ads during the 2018 Brazilian elections, this model found 835 ads (2.2%) that had not been correctly declared nor detected as political. Matias et al. [321] conducted an audit study on Facebook and Google's political ad policy enforcement through self-published ads, finding that Facebook applies their policies too restrictively, leading to 10 mistakenly prohibited ads (out of 238), while Google prohibited no ads. Sosnovik and Goga [456] compared platform, advertiser, and user perceptions of the definition of online political ads on Facebook through 63,400 crowdsourced ads labeled by volunteers. They found that social issue ads in particular see the highest error rate due to unclear policies, although users largely perceive them as political, and observe disagreement between automated classifiers for political ads trained on differently sourced sets of (non-)political ads. Moreover, several media reports have given anecdotal evidence of ads missed by Facebook's enforcement, both from politicians [153, 440, 484] and social issue organizations [440, 509]. Using the Ad Library, Cecere et al. [109] found

that COVID-19-related ads were more likely to be detected by Facebook, suggesting that these may have been falsely detected, and that ad policies were confusing to advertisers. Our study quantifies the performance of Facebook's enforcement on a larger and more representative scale than these previous studies, as we gather all *active* ads for *all* pages with at least one political ad, to analyze the prevalence of *both* false positives and false negatives.

From a transparency perspective, Edelson et al. [160] described and compared the efforts by Facebook, Google, and Twitter on a technical level. They later conducted a security analysis on transparency for Facebook's Ad Library in the U.S. [157], finding that adversarial political advertisers could evade transparency requirements through erroneous disclaimers and undisclosed coordinated behavior. We assess how advertisers may evade declaring their ads as political altogether, through which they also avoid transparency.

Further audits of Facebook's advertising platform found that advertisers can exploit ad targeting to infer private or sensitive user information [172, 203–205, 493], or to deploy highly targeted and biased ad campaigns [53, 198, 401], with users receiving inadequate targeting explanations from Facebook [53]. Facebook's ad delivery may also skew which users see which ads, potentially leading to discrimination based on gender or race [42, 241, 283], including for political ads [43].

7.3 Enforcement Errors and Their Impact

We introduce the two error types that affect the security of Facebook's political ad platform, i.e., the 'threat model,' and describe the actors that may either exploit these errors to induce harm, or that are themselves harmed by these errors.

First, ads may not be detected as political by Facebook, i.e., are *false negatives*. Once they are allowed to run, these missed ads harm the *integrity* of the online political ad ecosystem and of Facebook's transparency efforts. They result from Facebook failing to discover ads that advertisers did not properly declare, whether deliberately to avoid scrutiny or accidentally due to misinterpreting (ambiguous) policies [153, 440, 456, 509]. Malicious advertisers may have an incentive to not declare politically motivated ads, as this relieves them of the accompanying restrictions. They would not need to get authorized by Facebook (requiring identification) nor display who paid for the ad [161]. Moreover, users will be unaware that the advertiser is attempting to influence them as the ad interface will not reflect that the ad is political [161], and they might be shown the ad even if they requested to see fewer political ads [234]. The advertiser can then abuse these flaws to spread disinformation or prohibited content (e.g., voter suppression), or engage in 'coordinated inauthentic behavior' where accounts conspire to run influence campaigns [157, 409], without being publicly identified. Moreover, such an advertiser can circumvent bans on political ads, as was (temporarily) the case after the 2020 U.S. elections [3]. Finally, advertisers may want to evade transparency

and accountability: undetected ads disappear from the Ad Library once they become inactive, leaving researchers and journalists unable to discover policy-violating content or hold advertisers and Facebook accountable for compliance with and enforcement of the political ad policy.

Second, ads may be incorrectly detected as political by Facebook, i.e., are *false positives*. As detected ads are taken down and may even result in pages being restricted from running ads or being deleted, Facebook reduces the *availability* of legitimate advertisements through these errors, whether the ads concern social issues (but do not influence public opinion) or are purely commercial. On the one hand, these can result from Facebook applying their policies too restrictively or erroneously and detecting ads that are in fact not political. For example, Facebook's enforcement errors were found to hinder public health messages related to COVID-19 [109, 438] and vaccines [251, 342, 468]; were thought to unduly insinuate political division for social themes [192, 212, 343, 410, 448, 450, 516]; or resulted from false name matches [188, 317, 321, 322]. On the other hand, advertisers themselves may introduce false positives by (voluntarily) over-declaring ads that are not in scope of the political ad policy [456], possibly due to incorrectly or overly cautiously interpreting this policy or because they fear the ad will otherwise be erroneously detected and taken down. False detections also erode trust in enforcement as a whole, as they suggest that the automated decision models are unable to properly distinguish political ads, and further reduce the quality of input data to these models.

Throughout this paper, we label ads as follows:

Considered as political by				
advertiser	Facebook	us	Label	Type
✓	N/A	✓	declared	True positive
✓	N/A	✗	(over-)declared	False positive
✗	✓	✓	detected*	True positive
✗	✓	✗	(over-)detected*	False positive
✗	✗	✓	undetected*	False negative
✗	✗	✗	–	True negative
✗	(✓ OR ✓)		*undeclared	

7.4 Data Collection

To understand the dynamics and possible shortcomings of Facebook's ad policy enforcement, we must capture the full lifespan for all relevant ads. The Ad Library API is insufficient for this purpose: it only returns ads *once* they are known to be political, which crucially excludes the period *before* Facebook enforces upon an ad, and omits ads that Facebook *never* enforces upon altogether. We therefore develop a novel large-scale data collection pipeline using the Ad Library web portal, which lists all active ads for a

given page, regardless of whether they are political. In this section, we first define the scope of our study and present our data collection method. We then describe and validate the resulting data set, and discuss the ethics of our data collection as well as the impact of its limitations on our study.

7.4.1 Scope and method

Our data collection started on July 9, 2020, and was initialized with all pages that were present in any Ad Library Report for the past week since April 21, 2020, i.e., all pages that had published any political ads relatively recently. Until January 12, 2021, we continuously added pages newly appearing in the most recently available Ad Library Report. We consider the resulting set of pages as the scope of our study. We continued collecting ads for these pages for four more weeks, i.e., until February 9, 2021. This period therefore covered the pause on political ads in the United States after the elections on November 3, 2020 [3]. Our data collection covered all 71 countries with an Ad Library report at the time of our measurement. Section 7.G lists the dates when reports were first available and when we started tracking their pages.

For every page that is in scope, we scrape its ads from the Ad Library web portal [17] with a target period of 24 hours, as well as page metadata with a target period of 14 days. We request all currently active ads which had impressions in the previous 7 days in any country; we do not apply any other filter. Additionally, for every ad, we gather its contents and metadata 14 days after its first observation through the ad snapshot tool used in the Ad Library API. As this endpoint reports the ad's most recent state, even if already inactive, this allows us to observe any enforcement by Facebook within 14 days of the ad's publication. For an ad detected within 14 days, we assume that ad detection led to its deactivation. We then calculate the activity period of an ad as the time between its first and last (daily) observation, assuming that the ad was published just before the former and detected just after the latter, with a 24-hour margin due to our scraping frequency. Moreover, whenever we analyze the activity period of an ad, we require that we likely captured the full lifespan of an ad, and therefore exclude ads active during the final four weeks of our data collection (reducing right censoring) or before/during our first scrape for a given page (reducing left censoring). Finally, we retrieve all political ads that were active during our measurement period through the Ad Library API on March 30, 2021, in the subset of countries where API data was available to us (covering 80% of scraped ads, Section 7.G).

7.4.2 Data set description

In total, we observed 33.8 million unique ads during our measurement, of which 4.2 million were declared or detected as political (Table 7.2). As Facebook only provides ranges for ad spend and impressions for political ads, we estimate that these had around 100 billion impressions and cost around 1 to 1.4 billion U.S. dollars. We see that United

Table 7.2: Page/ad counts for the top 10 and other countries. For spend and impressions, we calculate the lower and upper bound based on the ranges available in our data.

				for political ads		
Country	# pages ▽	with ad	# ads	# ads	Spend (10^6 USD)	Impressions (10^9)
U.S.	90,018	69,815	21,934,716	1,902,473	810–1,146	45–53
Brazil	39,675	33,459	1,039,109	696,612	13–27	8.4–10
India	13,798	10,722	779,670	121,269	2.9–3.6	3.9–4.4
Italy	11,758	10,049	384,808	124,767	6.5–22	2.7–3.2
U.K.	10,558	8,016	2,456,921	96,660	17–33	2.4–2.8
Germany	10,223	8,501	782,078	121,482	13–30	2.8–3.2
Ukraine	8,632	7,437	315,425	133,006	3.2–17	3.4–4.0
Mexico	7,229	6,145	275,028	88,406	4.9–6.9	3.8–4.4
Canada	6,352	5,198	529,107	76,076	13–22	1.8–2.1
Romania	5,708	4,927	189,389	104,228	5.8–9.1	2.8–3.3
Other	61,873	50,761	5,142,518	726,382	54–109	18–21
Total	265,824	215,030	33,828,769	4,191,361	944–1,426	95–112

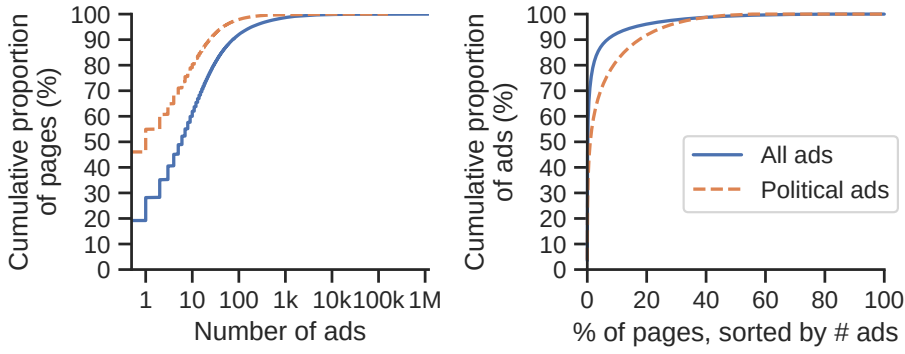


Figure 7.1: Distributions of observed ads and pages.

States advertisers dominate our data in terms of the number of ads placed overall, as well as in political ad count, spend, and impressions until the 2020 U.S. elections (Section 7.B), when Facebook restricted U.S. political ads [3].

We observed ads across 215,030 pages (by definition, all of these pages had at least one political ad ever); we never observed any ads for 50,794 additional pages in scope. Small advertisers represent the majority of these pages: the median page posts fewer than 6 ads and 1 political ad respectively (Figure 7.1). Conversely, a small percentage of pages account for the majority of ads: the top 20% advertisers posted 92.5% of all ads and 81.9% of all political ads. Distributions are similar between U.S. and non-U.S. advertisers. To conserve resources, we manually analyzed the largest advertisers up to then on four occasions (Section 7.B), and discarded those that we considered unlikely to intentionally publish political ads. Before discarding, these advertisers placed 5,862,808 ads (17.3% of all ads). Our analysis in Section 7.C confirms that these pages minimally placed political ads (0.011% of their observed ads), supporting our decision to discard them.

7.4.3 Data set validation

We assess the coverage of our data set both internally and externally to determine how often we were unable to retrieve all available ads. As an internal validation, we compare the expected number of available ads included in Facebook portal data to the number of actually observed ads. As an external validation, we compare the observed political ads with those retrieved from the Ad Library API.

We summarize our coverage in Figure 7.2. We missed 19.8% of ad observations, most often due to a limitation in Facebook's systems: even though the portal is able to report that a page has over 50,000 ads, it fails to actually load more than 7,800 ads per scrape. Very large advertisers therefore bear the bulk of missing observations. We also miss the first ads from newly seen advertisers due to a delay of usually three days between a page's first political ad impression and its appearance in the Ad Library Report [149]. Otherwise, discrepancies are due to our scraping frequency or setup: we miss ads that disappear during a scrape, that only appear between scrapes or after the last scrape, or when the scraper (partially) failed. For 7.5% of ads, we could not make the 14-day snapshot: this is largely due to resources being unavailable or restricted through Facebook's snapshot tool, or because the 14-day interval was outside our measurement period. Finally, for 184 pages (0.07%), we failed to retrieve page metadata.

Based on the maximum number of missed observations per page, we estimate to have missed at least an additional 6.4% of ads. Weighted by the observed proportion of political ads per page, we estimate to have missed 11.4% of political ad observations. We note that the (unknown) number of unique ads that we were unable to retrieve is significantly lower than the number of missed ad observations, since many ads are active for more than one day. Finally, based on the API data, we missed around 1 million political ads (24.8%), with an estimated combined spend between 188 and 550 million USD, and 24 to 31 billion impressions. While missing data may introduce risks to research validity [191],

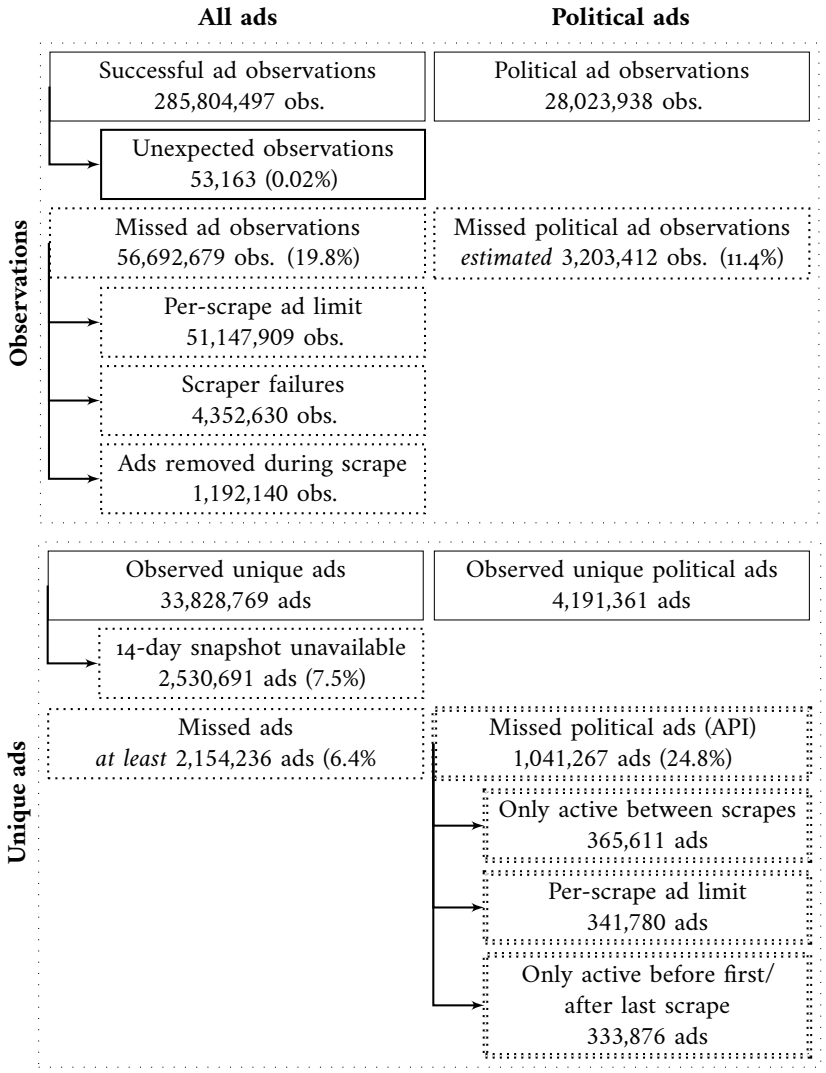


Figure 7.2: Summary of data set coverage.

our findings are lower bounds mainly calculated in the aggregate, which are less affected by our data gaps. We therefore believe that our data and results remain sufficiently representative for the Facebook political ad ecosystem.

7.4.4 Ethics

We follow ethical guidelines for Internet measurement and cybersecurity research [69, 218, 313]. Our data collection does not affect any non-advertiser Facebook users and we do not observe any personally identifiable information on them. Our research received an IRB exemption as it does not involve human subjects. As part of publicly available metadata, we collect the disclaimer that the advertiser provides to Facebook for the ad authorization process [197]; in the case of individuals, this may include personally identifiable information such as their legal name. We only process this data in the aggregate and do not use it to identify any individual. Similarly, we do not name any specific advertiser to avoid inflicting harm resulting from flaws in Facebook's enforcement.

We collect only publicly accessible data. As Facebook states that "more than 2 million people visit the Ad Library every month" [6], we do not expect this data collection to significantly affect the availability of the Ad Library, and we did not observe any service outage possibly caused by our scraping. While Facebook's 'Automated Data Collection Terms' [63] may prohibit scraping, we believe that our research is in the public interest, and that its societal benefits justify the technical resources consumed from Facebook, as well as potential reputational and financial harm on Facebook. Institutions, civil society organizations, and researchers have previously called for improved transparency for all ads [5, 120, 156, 164, 235, 250, 252, 253, 295, 404, 497], confirming the value of our data set. We share our data and methods with other researchers at `https://osf.io/7tw3e/`. In the context of prior work, we already communicated with Facebook to discuss their ad review and transparency, and presented to them the overall issues and recommendations that we also analyze in this paper.

7.4.5 Limitations

The definition of our scope leads us to only track pages with at least one known political ad (declared or detected) as recorded in the Ad Library Report. If a page never gets caught or is exempt (news pages [16]), we will therefore not track its ads, potentially missing their false negatives. Likewise, we cannot cover advertisers in countries where declaration is voluntary, as Facebook does not publish an Ad Library Report there [11].

As quantified in Section 7.4.3, we do not achieve full coverage of the ads published by advertisers within our scope. On the one hand, coverage is affected by the trade-off between limited resources on our side and the large number of pages in scope. The 24-hour scraping period means we miss ads that were active only between two scrapes, and limits the granularity of observed activity periods. Our snapshot delay leads us to miss ad status changes beyond 14 days. However, we consider any changes after 14 days less likely to be due to Facebook's own enforcement system,[6] but rather due to external

[6] Facebook started its pause on political ads 7 days before the 2020 U.S. election, a.o., to "re-review[ads] for policy violations" [481].

reporting. We also request ads shown in any country; Facebook provides a filter by country, but this would prohibitively multiply the required resources.[7] We therefore assign pages and their ads to a country based on a heuristic, i.e., top spend on political ads. On the other hand, delays, flaws and changes in Facebook's systems further reduce coverage. Delays in the publication of the Ad Library Report [149] and a limit on the number of retrievable ads cause us to partially miss ads from newly added and very large advertisers, respectively. We also experience infrequent failures of our scrapers, due to changed request methods or unavailable resources, or race conditions during one scrape (e.g., leading to duplicate ads). Beyond these unobserved ads in scope, we do not know the total number of ads on Facebook, which prevents us from quantifying true negatives and calculating metrics that depend on it. However, we expect true negatives to be much more prevalent, and therefore select classification metrics that are more robust against this class imbalance.

Finally, limitations result from Facebook's transparency implementation. Without full metadata on all ads, we cannot quantify the impact in terms of spend and impressions of undetected political ads. We also have no visibility into ads that are caught during initial review and are therefore prevented from running altogether. More abstractly, we rely on Facebook's Ad Library functioning properly, i.e., returning the actual, complete set of (non-)political ads from all pages [83, 497, 512]. While we have no reason to believe this is not the case, we also have no way to confirm this for our data set, due to a lack of transparency into Facebook's architecture. Crowdsourced ads may allow to audit the accuracy of the Ad Library, albeit not completely [83, 134]. Moreover, while we are the first to conduct large-scale data collection through the web portal, researchers and organizations have documented consistency, completeness, and reliability issues with the Ad Library API [92, 118, 149, 157, 164, 168, 411, 434]. These challenges in comprehensively obtaining all currently active ads ultimately harm Facebook's transparency efforts.

7.5 Ad-level Enforcement

We first examine enforcement of individual ads, independent from the advertiser. We start by quantifying the prevalence of enforcement, that is, how frequently Facebook is taking down ads for not having the required disclaimers, and determine the exposure that violating ads had before detection. We then survey how often an enforcement decision made by Facebook is appropriate, especially with regard to ads that likely should not have been taken down (*false positives*) and where incorrect enforcement harmed advertisers.

[7] However, we find that this filter can also be unreliable, with some ads only being available when no country filter is set.

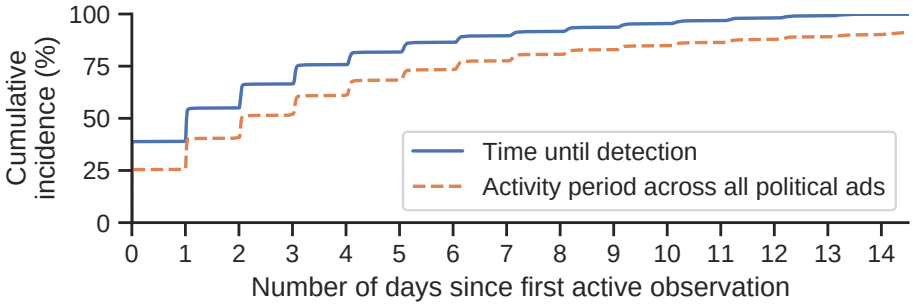


Figure 7.3: Activity period for detected and all political ads where we likely observed the full lifespan (Section 7.4.1) and observed detections within 14 days.

7.5.1 Current ad policy enforcement

Within our measurement data, 72,678 ads were marked at some point as 'detected,' i.e., political but not properly declared, within 14 days after the ad's first activity. These detected ads therefore represent a minor share (1.7%) of all 4.2 million observed political ads. Edelson et al. [157] reported a 9.7% detection rate for May 2018–June 2019, suggesting policy awareness and compliance may have since improved. Moreover, Facebook has stated that "between March 1[, 2021] and Election Day, [they] rejected about 3.3 million ad submissions that targeted the US without completing the authorization process before they could run" [6], suggesting that Facebook's initial ad review already catches most violations, although the lack of detail makes a reliable comparison difficult (e.g., authorization may be subject to separate review, and an advertiser could try and resubmit an ad until it passes).

Next, we analyze whether Facebook prevents violating ads from gaining much exposure by measuring how quickly Facebook takes down an ad that should have been declared as political. Detection of ads that slipped through the initial ad review is relatively fast (Figure 7.3): 40% of ads were detected within less than 1 day, with the median activity period being less than 2 days. Detected ads are also active for shorter periods than any political ad, for which the median activity period is less than 3 days, suggesting that ads are primarily detected while they are still active. However, violating ads may still enjoy significant exposure in budget, impressions, and time. We estimate the detected ads to have accumulated spending between 12.2 and 20.7 million U.S. dollars and between 2.1 and 2.4 billion impressions, i.e., instances where a user saw the ad without the proper context that it was political. 5,885 ads (8.1%) were active for over a week, meaning detection occurred very late. Moreover, we find 49,263 ads that were likely detected only after they became inactive, as they became inactive within 14 days, were not yet marked as political after 14 days, but were present in the Ad Library API. The advertiser could therefore display their violating ad for the desired duration. These 55,148 detections combined do not prevent most or any user harm, as most or all intended ad impressions still occur. Instead, they are only useful for the secondary goal of transparency (as the ads are then

included in the Ad Library) and for any potential disciplinary measures taken against the page.

7.5.2 Ads incorrectly detected as political

When Facebook takes down ads for a lack of disclosure of their political nature, some of these decisions are incorrect, i.e., *false positives*. For example, the takedown of 1,413 ads (1.16% of all detected ads) was later undone, possibly after an appeal by the advertiser. These reflect admissions by Facebook that the ads were false positives and should not have been disabled. To study false positives more systematically, we labeled a randomly selected sample of 300 correctly declared and 300 detected ads. We restricted these samples to advertisers in the United States to ensure that annotators could interpret ad text and context. Three authors determined whether each ad was within or outside the scope of Facebook's political ad policy. They were instructed to adhere as closely as possible to Facebook's definition, i.e., not apply their own interpretation of what should be a political ad. In a subsequent meeting, the annotators revisited disagreed-upon ads, and reassigned a final agreed-upon label in the case of simple labeling errors (agreement on the definition, but for example a missed reference to a politician). Otherwise, if they considered Facebook's definition too ambiguous, in particular on whether an ad sought to influence public opinion, they recorded "disagreement" as the final outcome. Using Krippendorff's α [275], we achieve an inter-rater reliability of 0.81, i.e., sufficiently strong agreement for reliable conclusions.

Table 7.3 lists the results of our labeling. For *declared* ads, a false positive indicates that an advertiser unnecessarily declared that ad. Across our sample, we observe 3.3% over-declared ads, suggesting the practice is rare. 80% of declared ads are related to politics and elections, which are clearly in scope of Facebook's ad policy and should therefore be declared. Across all observed ads, Facebook does not appear to retroactively mark declared ads as non-political, i.e., Facebook does not (need to) check whether a declared ad falls within the scope of its policy. For *detected* ads, a false positive indicates over-enforcement by Facebook, unduly taking down the ad. Across our sample, a majority of detected ads (55%) should not have been enforced upon; if we extrapolate this rate to all detected ads, 67,433 ads should not have been taken down. This suggests that Facebook's ad detection is overly aggressive. Edelson et al. [157] observed a 79% false positive rate through a similar manual analysis, corroborating our finding that this rate may be very high.

The annotators also described the ad topic, using an inductively developed codebook that was aligned in their meeting (Section 7.E). 24% of false positive detected ads concern commercial products or services, and it was not immediately obvious why Facebook detected those ads as political. Among ads where the likely cause of error was more discernible, most errors concerned COVID-19-related and health ads, for which the ambiguity in the definition of 'social issues' ("seek to influence public opinion") makes confusion for advertisers and Facebook's review more understandable. It appears that these policy ambiguities account for most errors; we attribute only 6 false positives to a

Table 7.3: Manual categorization of 300 declared and 300 detected ads, grouped by annotators' assessment of whether these are political per Facebook's ad policy. ⊙ : Related to social issues. Percentages are given within the sets of declared and detected ads, respectively; the margin of error is for a 95% binomial proportion confidence interval.

Ads considered **political** (*true positives*)

	declared		detected	
Topic	#	%	#	%
By a political figure/organization	143	47.7	1	0.33
About a political figure/organization	61	20.3	15	5.00
About elections	35	11.7	13	4.33
Political Values and Governance ⊙	10	3.33	10	3.33
Civil rights ⊙	5	1.67	13	4.33
Environment ⊙	6	2.00	4	1.33
Economy ⊙	6	2.00	3	1.00
Other ⊙	9	3.00	6	2.00
Total (Precision)	275	91.7	65	21.7
Margin of error		±3.1		±4.7

Ads considered **non-political** (*false positives*)

	declared		detected	
Topic	#	%	#	%
Commercial product/service	0	0.00	73	24.3
COVID-19-related	0	0.00	24	8.00
Health ⊙	5	1.67	18	6.00
News/media	1	0.33	8	2.67
Apps/games/websites	0	0.00	8	2.67
Arts/Personalities	0	0.00	7	2.33
Other	0	0.00	24	8.00
Other ⊙	4	1.33	4	1.33
Total (False discovery rate)	10	3.33	166	55.3
Margin of error		±2.0		±5.6

Ads **disagreed** upon by labelers

	declared		detected	
Topic	#	%	#	%
Environment ⊙	4	1.33	13	4.33
Insignificant reference to politics	2	0.67	9	3.00
Food assistance ⊙	2	0.67	6	2.00
News/media	0	0.00	8	2.67
COVID-19-related	1	0.33	6	2.00
Health ⊙	2	0.67	4	1.33
Other	1	0.33	3	1.00
Other ⊙	3	1.00	20	6.67
Total (Disagreement rate)	15	5.00	69	23.0
Margin of error		±2.5		±4.8

likely false "keyword" match (e.g., a shop called 'Mayors'). Matias et al. [321] observed a 4.2% false positive rate across ambiguous issue ads and false matches. Our findings suggest that in addition, commercial ads, where the reason for detection is less clear, represent a significant share of false positives. Finally, the annotators could not agree on the label for 5% of declared and 23% of detected ads, highlighting the difficulty of interpreting Facebook's ad policy consistently. This disagreement mainly involved ads relating to social issues, where it was unclear whether the ad sought to influence public opinion, or ads with incidental references to politics, such as a candidate's yard sign being visible in a real estate listing.

Summary We find that detected political ads without a disclaimer account for only a small share (1.7%) of observed political ads, and that Facebook detected them rather quickly. Still, we see 55,148 detected ads running for more than a week or for their full intended duration, making detection largely ineffective at preventing users from seeing these violating political ads. In addition, detection appears to be very imprecise: we find that 55% of detected ads in the U.S. should not have been taken down (*false positives*), harming advertisers by making their legitimate ads unavailable to users.

7.6 Page-level Enforcement

We continue at the level of a Facebook page to examine whether enforcement appears to be consistently and correctly applied across all ads of a page. We start by describing how advertisers react to takedowns of a page's ads. We then classify pages to describe the current composition of Facebook's Ad Library based on a page's likely political intent, and identify likely *false negatives* as ads published by pages with a clear political purpose that Facebook failed to detect.

7.6.1 Reaction to enforcement

We first analyze whether advertisers are able to repeatedly violate Facebook's ad policy. We observe detections of undeclared political ads for 13,900 pages (5.2%), again lower than the 68.3% rate observed by Edelson et al. [157] for May 2018–June 2019. Overall counts of detected ads per page were low (Figure 7.4, left). However, 7,535 pages (54.2%) did not declare any political ads, so their only political ads were those that were detected (Figure 7.4, top), suggesting they may have unintentionally posted ads that were deemed to be political, possibly due to insufficient awareness of or ambiguity in ad policies. No advertisers appear to have placed many and mostly political ads without declaring them (Figure 7.4, top right); high absolute counts of detected ads are an artifact of an overall high volume of ads (Figure 7.4, bottom right). It does not appear that Facebook frequently banned pages as an enforcement action after detecting undeclared ads: only 458 pages were deleted some time after an ad detection, or 3.3% of pages with detected undeclared ads. (For reference, 7.2% of all pages in scope were ever deleted.) 373 of these 458 pages

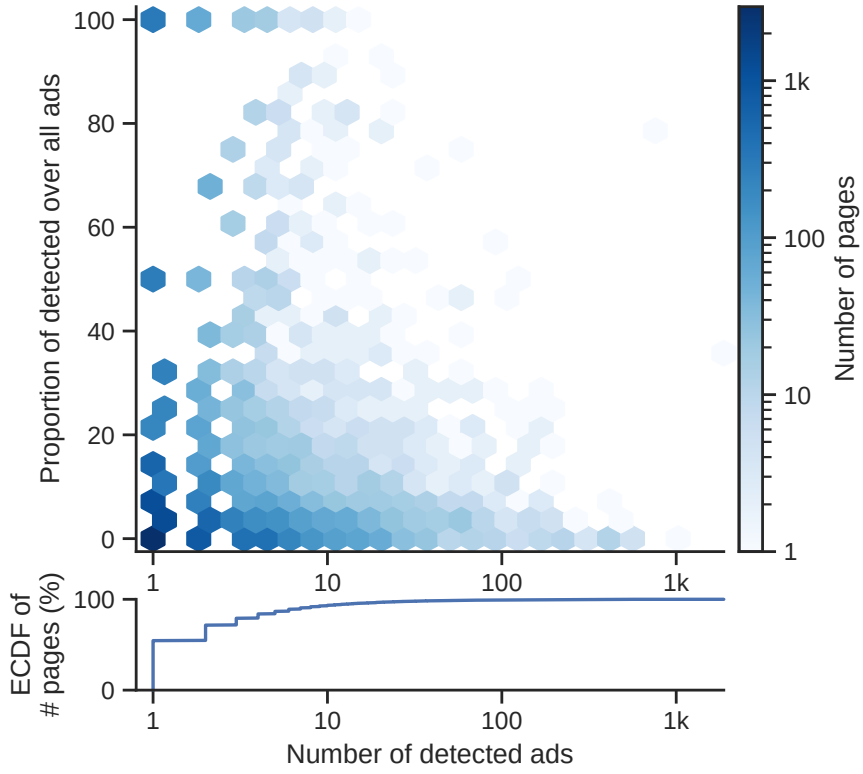


Figure 7.4: Detected ads versus their share of all ads for a page.

even continued placing new ads between their last taken down ad and deletion of the page.

Next, we analyze whether advertisers declare more ads after ads have been taken down, i.e., whether Facebook's enforcement increased adherence to its ad policies. Ideally, this reaction should prevent future violations and protect users from being exposed to unmarked political ads. Out of the top 75 pages ranked by detected ad count (listed by class in Figure 7.5), 22 increased their proportion of ads declared as political after being detected (①–④): 5 started declaring continuously (①) and 5 others only shortly did not declare (②). However, increased declaration was only short-lived for 12 pages (③–④). Furthermore, 48 pages (⑤–⑥) barely declared any ad as political and often had a steady stream of violating ads taken down by Facebook. (This includes 39 news aggregator pages (⑥) that are likely not exempt from declaration, unlike more traditional news organizations [16].) This suggests that the most frequent offenders (in absolute terms) did not face any durable restriction in their ability to run ads as a potential disciplinary measure imposed by Facebook to increase compliance. Despite sometimes frequent and prolonged violations, all pages in the top 75 continued publishing ads after detection.

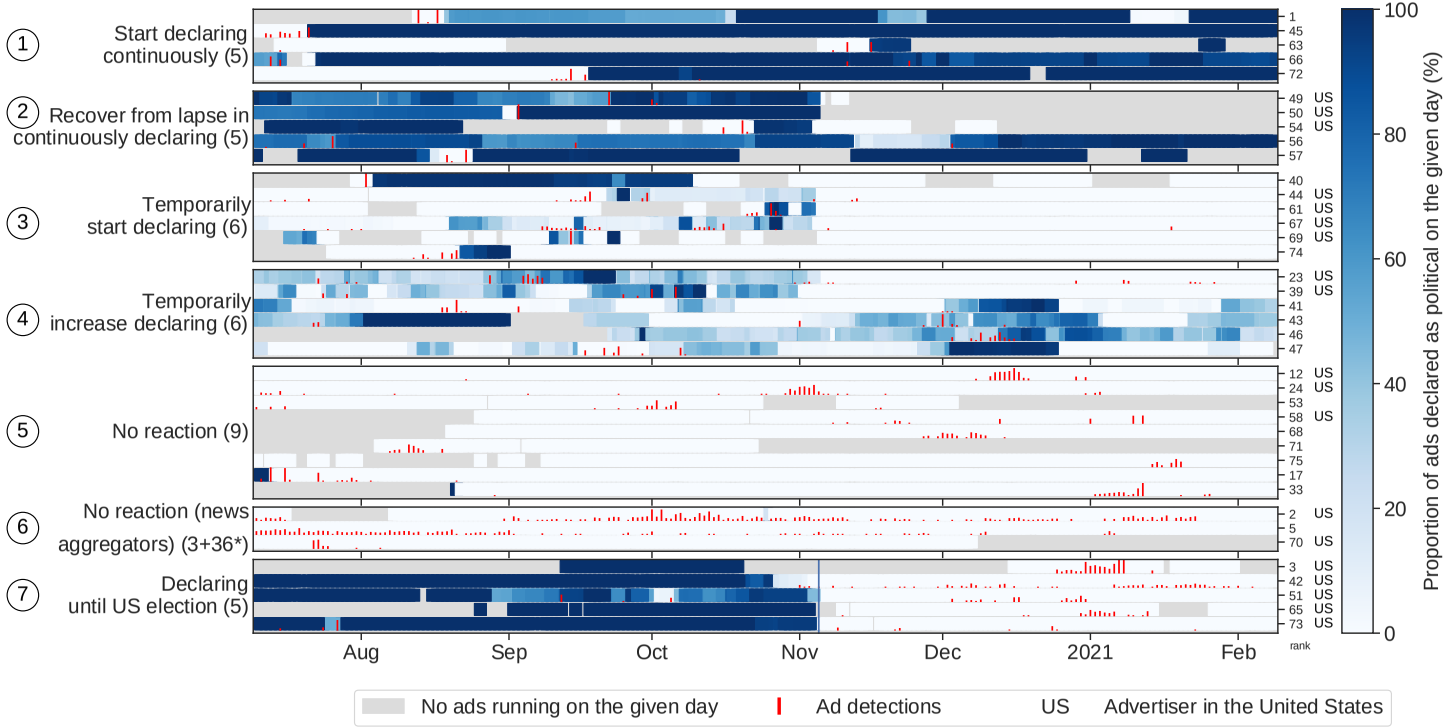


Figure 7.5: Reactions to ad detection by the 75 advertisers with most detected ads. Each row corresponds to an advertiser, with shades of blue indicating the proportion of ads the respective advertiser declared as political over time (aggregated daily). Red markers at the bottom of each row denote when ads were detected and taken down by Facebook. Advertisers are grouped into seven classes (denoted by ⑦) based on their reaction to enforcement (start/increase/resume declaring political ads) and the duration of increased declaration (continuously/temporarily/no reaction). For example, group ⑥ contains advertisers that have a steady stream of ads taken down by Facebook but do not show any apparent reaction to enforcement (they continue not declaring any ads as political). These advertisers are news aggregator pages; we show three examples and omit 36 similar pages (*).

After the 2020 U.S. election, Facebook temporarily disallowed political ads on their platform [3]. Instead of ceasing to advertise, 5 of the top 75 violating pages continued running ads but *stopped* declaring them as political (⑦), even though they were clearly of political nature as they had previously declared (nearly) all their ads as political. Even though these 5 pages advertised merchandise such as T-shirts with political messages, or were advocating for civil rights and environmental policy, Facebook only detected and took down 3% of their ads running after the election. Overall, 1,018 pages ran 71,426 undeclared ads after the U.S. election, whereas they only ran political ads before then. These pages did not appear to be deterred by the political ad pause, and Facebook did not effectively prevent them from running ads that were very likely political. This failure of enforcement rendered the ad pause less effective, and put these violating pages at an unfair advantage over advertisers who did comply and ceased running political ads as required [284, 435].

7.6.2 Current enforcement by advertiser class

Facebook's policy on ads that require disclosure is broader than just ads published by obvious political actors such as parties or candidates [252]. Consequently, the Ad Library Report also lists advertisers beyond those actors, such as those placing ads about social issues, in partnership with (on behalf of) political actors, or with (non-partisan) calls to vote, next to advertisers with incorrectly detected ads (Section 7.5.2).

Page classification

To quantify the prevalence of different types of advertisers, we match internal and external data sources with observed Facebook page metadata to classify pages into one of four topics. If a page is listed in multiple sources, we select its main class using the following order: (1) political, (2) government, (3) media, and (4) issue-related.

- We retrieve political committees for the 2020 U.S. elections from Facebook's Ad Spending Tracker [480] (matched on page name) and the OpenSecrets project of the Center for Responsive Politics [368] (matched on page ID).
- We retrieve political candidates and parties registered for the 2020 Brazilian municipal elections from the Superior Electoral Court [446] (matched on CNPJ).
- We retrieve pages who identified themselves during Facebook's advertiser authorization process through a Federal Election Commission identification number (FEC ID) for U.S. political pages or U.S. government credentials for U.S. government pages from our data (matched on page ID).
- We retrieve media organizations from Media Bias/Fact Check [324] and NewsGuard [360] (matched on page alias), with local pages for large news aggregators manually added.

- We retrieve U.S. nonprofit (tax-exempt) organizations as potential issue-related pages from National Center for Charitable Statistics [80, 357] and Internal Revenue Service [167] data (matched on disclaimer and ZIP code).
- We retrieve manually curated *Explore* lists [386] containing political, government, media, and issue-related pages from CrowdTangle [131] (matched on page ID).
- We enumerate the most common Facebook page categories for pages within the previous data sets, manually select those categories that are sufficiently specific to a class, and then retrieve all pages within those categories (Section 7.F) from our data (matched on page ID).
- We retrieve all pages that completed Facebook's advertiser authorization process from our data (matched on page ID).

Overall, 59.7% of pages fall into one of the four categories. We distinguish an additional 11.3% of pages outside these topics that completed the authorization process, as this implies a genuine intent to at least sometimes place political ads.

Distribution of political ads over classes

We first analyze the composition of advertisers listed in the Ad Library Report globally, i.e., within our scope of pages with at least one recent (declared or detected) political ad (Section 7.4.1). We expect mostly political and issue pages to appear in the Report; however, based on our classification (Table 7.4), 'obvious' political advertisers only represent 39% of measured pages, with a further 8% that are issue advertisers; these combined account for 73% of observed political ads. Only 53% of all pages were authorized to declare ads as political. Unauthorized pages may have no political motive, suggesting they inadvertently published ads that fall under the political ad policy,[8] or their ads were incorrectly detected as political by Facebook. Alternatively, they may be political actors that refuse to authorize themselves, or may be unable to do so due to Facebook's policies, e.g., if they are outside the country in which they want to run political ads [157].

We further analyze whether certain page classes are more likely to have such "unintentionally" undeclared and detected ads by comparing *detected* with *overall* ad counts (Table 7.4). Government (6% detected vs. 4% overall) and issue (16% vs. 8%) pages are overrepresented, hinting at discrepancies between their and Facebook's understanding of which ads should be declared. Media pages account for the most detected ads in absolute numbers (34%), but place ads in similarly high volumes (35% of all ads). 46% of authorized pages and 21% of political pages failed to declare at least one ad that was later detected, even though Facebook's ad policy requires all ads from or on behalf of political figures to be disclosed.

Next, we measure the proportions of political ads over all ads per page and class. If an advertiser is political in nature, we expect them to have 100% political ads, either because they properly declare all their ads or because Facebook detects their undeclared ads.

[8] These ads then likely triggered inclusion in the Report; due to delays in this inclusion, we cannot observe these one-off political ads (Section 7.4.5).

Table 7.4: Distribution of observed (political) ads over identified advertiser classes. Classes do not overlap.

	Pages		All ads		Political ads	
Type of page	#	%	#	%	#	%
Political	102,617	38.6	2,593,727	7.7	2,476,764	59.1
Government	9,686	3.6	781,102	2.3	190,057	4.5
Issue	20,250	7.6	1,722,973	5.1	585,326	14.0
Media	26,203	9.9	11,814,593	34.9	318,939	7.6
Other with authorization	30,108	11.3	1,954,851	5.8	604,897	14.4
Other with detected political ads	3,619	1.4	6,540,777	19.3	15,378	0.4
Other with only observed non-political ads	47,248	17.8	8,420,746	24.9	0	0.0
Other without observed ads	26,093	9.8	0	0.0	0	0.0
All with authorization	139,861	52.6	8,132,939	24.0	4,151,486	99.0
All with declared or detected political ads	143,657	54.0	21,511,759	63.6	4,191,361	100.0
All pages	265,824	100.0	33,828,769	100.0	4,191,361	100.0

	Political ad proportion (%)				% pages with	Detected ads		Detected pages	
Type of page	overall	Q_{1-3} per page			100% pol. ads	#	%	#	%
Political	95.5	100.0	100.0	100.0	80.8	6,010	8.3	2,965	21.3
Government	24.3	0.0	20.0	92.9	22.7	4,577	6.3	956	6.9
Issue	34.0	0.0	25.0	93.9	22.3	11,245	15.5	1,870	13.5
Media	2.7	0.0	0.0	50.0	12.2	24,395	33.6	2,858	20.6
Other with authorization	30.9	44.4	100.0	100.0	52.9	11,073	15.2	1,632	11.7
Other with detected political ads	0.2	0.8	3.8	14.3	4.9	15,378	21.2	3,619	26.0
Other with only observed non-political ads	0.0	0.0	0.0	0.0	0.0	0	0.0	0	0.0
Other without observed ads	0.0	– – –	– – –	– – –	0.0	0	0.0	0	0.0
All with authorization	51.0	89.5	100.0	100.0	69.4	32,803	45.1	6,412	46.1
All with declared or detected political ads	19.5	80.0	100.0	100.0	66.3	72,678	100.0	13,900	100.0
All pages	12.4	0.0	80.3	100.0	44.3	72,678	100.0	13,900	100.0

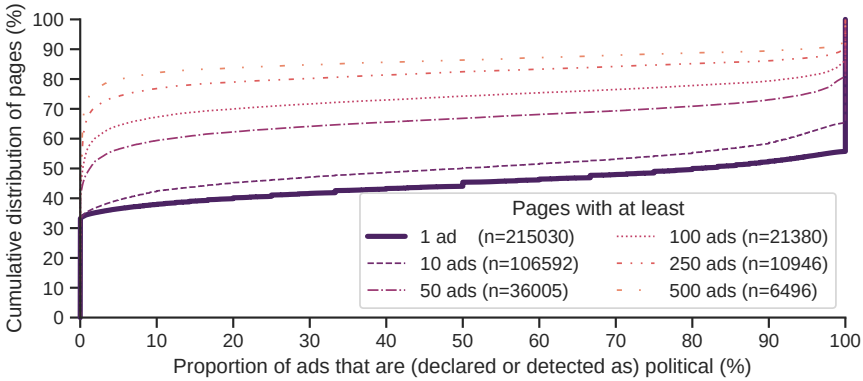


Figure 7.6: Distribution of political ad proportion over pages.

Indeed, this largely holds for identified political pages (Table 7.4), where 81% had only political ads. However, for government, issue, and media pages, this share is much lower, at 23%, 22%, and 12% respectively, and median proportions of political ads of 20%, 25%, and 0% respectively, showing that the Report contains many (classes of) pages whose ads are mostly non-political. Across all pages with observed ads (Figure 7.6), we similarly see that 44% had only political ads, while 33% had no political ads observed during our measurement, with the latter increasing for larger advertisers, suggesting these may not have political intent.

Overall, we find that 47% of pages in the Ad Library Report are not authorized, with over 33% of pages hardly publishing political ads over time. These pages may have had incorrectly detected ads (*false positives*) or placed an ad that they did not consider political even though Facebook did. Indeed, government and issue pages are much more likely to have ads detected by Facebook. When pages have no clear political motive, the advertisers as well as Facebook must determine at the individual ad level whether the ad is in scope of the political ad policy, which may be more prone to interpretation errors and disagreements and therefore lead to enforcement errors. Conversely, we consider 39% of pages to be core political actors. Although Facebook's policies require any ad made *by* a political actor to be declared [23], enforcement is still necessary as 21% of such pages have at least one detected ad. Next, we analyze whether these pages had any ads that were neither declared nor detected by Facebook.

7.6.3 Missed ads by political advertisers

Through our classification from Section 7.6.2, we can identify advertisers that are known or self-declare to be political actors. For these advertisers, Facebook's ad policy explicitly mandates that *all* their ads be declared ("ads *made by*" a political actor). If these advertisers fail to disclose all their ads, Facebook should detect them. With this premise, we can measure whether those pages had any undetected ads that Facebook's enforcement missed

Table 7.5: Ads from clear political advertisers missed by Facebook's enforcement (false negatives). Groups may overlap.

		Undetected ads		Pages with ≥1 undetected ad		Detected ads	
Political page list	Country	#	%*	#	%	#	%§
Ad Spending Tracker	US	33	0.01	16	1.93	1	0.00
FEC-registered organizations	US	1,035	0.17	42	2.76	40	0.01
OpenSecrets.org committees	US	129	0.03	24	1.74	14	0.00
CNPJ-registered entities	BR	7,607	1.57	2,038	11.99	668	0.14
CrowdTangle Explore lists	Int.	14,263	2.15	649	10.46	306	0.05
Facebook Page categories	Int.	103,808	4.33	16,116	16.53	5,764	0.25
All political pages		116,963	4.51	16,875	16.44	6,010	0.24
U.S.-only	US	10,940	0.96	1,187	4.75	416	0.04

* False negative rate ($FN/(FN+TP)$) § Proportion of ads detected by Facebook across all ads labeled as political by Facebook within the respective group of pages (TP).

within 14 days after the ad's first activity. By taking a more holistic approach where we identify groups of pages and ads where enforcement is required, we avoid introducing any interpretation of our own of what should be a 'political ad' [456], as well as errors from machine learning models that detect political ads [154, 439]. Instead, by selecting pages that we believe to be clearly in scope of Facebook's political ad policy, we have greater confidence that we observe genuine errors in Facebook's enforcement.

Table 7.5 shows the different lists of political pages that we derived from the external data sources. To ensure the precision of these lists, one author manually verified all pages with undetected ads from the three U.S.-based lists and removed entries that were not political actors. One such example was a media page that ran a few one-off ads on behalf of a presidential campaign and disclosed them using the campaign's FEC ID. For the larger list of advertisers that self-declare a political Facebook page category, such as 'Politician' or 'Political Organization,' we randomly sampled 50 pages from the U.S. (to ensure interpretability) for which one author manually confirmed that all 50 pages belonged to core political actors. We conclude that these lists are a reliable source of political pages that are required under Facebook's policy to declare all of their ads as political.

Across federal and state-level U.S. political advertisers, i.e., those that Facebook included in its Spending Tracker or verified through an FEC ID as well as major committees tracked by OpenSecrets, performance is nearly perfect with at most 0.17% missed ads, depending on the list (Table 7.5). If we broaden our view to include additional core political actors in the U.S., identified based on manually curated political Explore lists from CrowdTangle (a subsidiary of Facebook) or the category of their Facebook pages, Facebook misses more ads (0.96% in total). As these additional lists also cover advertisers for local elections, the increase in missed ads suggests that Facebook is less successful at identifying smaller

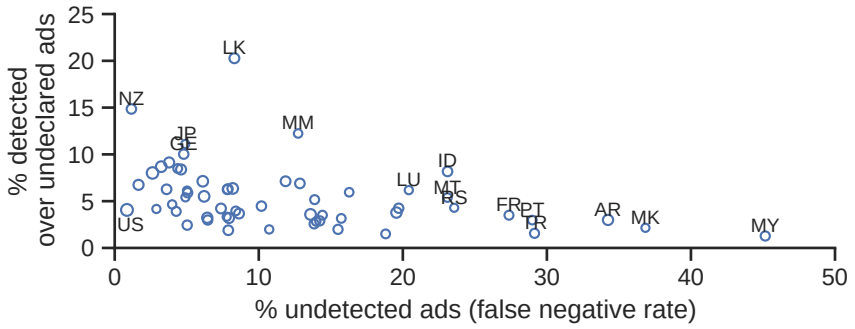


Figure 7.7: Rates of undetected and detected ads for countries with at least 0.01% of all observed political ads, across pages in political Facebook page categories.

political actors. In absolute terms, out of the combined 11,356 ads that U.S. core political actors failed to declare, Facebook was only able to detect for 416 (3.7%) that they were political and therefore enforce its ad policy. Facebook failed to detect the remaining 10,940 electoral ads, including ads from a U.S. senator and former presidential primary candidate with almost 10 million dollar in ad spend, who was able to run undisclosed (and undetected) ads after the U.S. elections.

We now analyze whether Facebook's enforcement is consistent globally through international lists of political actors, comparing performance in particular to that in the United States. For candidates and parties registered in the 2020 Brazilian municipal elections [446], 1.57% of ads went undetected, even though Facebook could easily match their provided CNPJ ID with the official list of political advertisers. Silva et al. [439] observed a slightly higher false negative rate of 2.2% in the 2018 Brazilian elections, albeit across all advertiser types. Across global lists of political actors curated by CrowdTangle, 2.15% of ads were missed. Across pages that classify themselves in a political Facebook page category, 4.3% of ads were neither declared nor detected. (Note that this is a lower bound, as pages could select generic categories that we do not include here. Within the other lists of political pages, we find 3,311 'Public Figure' and 715 'Personal Blog' pages, for example.) All of these false negative rates are worse than even the most broadly defined set of political actors in the United States. Across all political advertisers worldwide, we find a false negative rate of 4.5%, almost five times that of the U.S. 16.4% of pages published at least one political ad that was not detected, more than three times as many as in the U.S., even though they are obvious political actors.

We further break down rates of undetected ads for advertisers in political Facebook page categories by country (Figure 7.7), as these advertisers cover all countries in our scope. Again, Facebook misses the fewest ads in the United States (0.85% false negatives), whereas enforcement can be considerably worse in other countries: in absolute terms, Argentina, Brazil, and India have over 10,000 undetected ads each, while in relative terms, Argentina, North Macedonia, and Myanmar have between 30% and 45% undetected ads. This further

suggests that Facebook does not succeed in enforcing its policies consistently worldwide, leaving some users more exposed to violating political ads. This inconsistency may be due to language-specific model deficiencies [425], in particular if little training data is available [358]. Additionally, the review model may fail to incorporate different cultural contexts, with certain topics being considered politically sensitive only in some countries, as Facebook itself recognizes [365]. However, other confounding factors exist, such as the importance of individual pages, the reach of ads, or heightened attention due to ongoing elections [243]. These prevent us from reliably attributing performance differences to one or more causes.

Overall, of all 122,973 ads that political actors do not declare, Facebook only succeeds in detecting a very minor share of 6,010 ads or 4.9%, while failing to detect 116,963 ads, meaning that Facebook's enforcement is ineffective at discovering violating ads even from advertisers with a clear political intent. Moreover, if an advertiser properly declares their ad, Facebook does not have to make an enforcement decision, increasing the significance of any error made for an undeclared ad. In terms of potential exposure, the activity period for these undetected ads is similar to that for detected ads (Section 7.5.1). While Facebook does not disclose spend and impressions metadata for ads not known to be political, we use the average ad spend and impressions across each page's political ads to estimate that advertisers likely spent between 4.6 and 9.2 million U.S. dollars on undetected ads, and that these ads had between 2.2 and 2.4 billion impressions. As with detected ads, users are therefore exposed in great quantities to these violating ads. However, undetected ads lack even the transparency that comes with after-the-fact detection, since these ads disappear from the Ad Library and can no longer be scrutinized once they become inactive.

Summary Detection led 22 of the 75 pages with the most detected ads to correctly declare more political ads. However, the remaining pages can and do continue publishing undeclared political ads even after Facebook detected their undeclared ads, and after Facebook prohibited political ads in the U.S. Facebook's policies appear to result in non-political pages being listed in the Ad Library Report as having political ads, which may have been inadvertently published or erroneously detected. Conversely, we consider only 39% of pages to be core political actors, who we expect to declare all their ads or else get detected by Facebook. Unfortunately, we find at least 116,963 ads from these clearly political advertisers that were missed by Facebook's detection (*false negatives*). Moreover, these missed ads are unevenly distributed worldwide: while for U.S. advertisers only 0.85% of ads are missed, we see a false negative rate of up to 45% in other countries. Put differently, for political pages only 4.9% of undeclared ads are detected, resulting in at least an estimated 2.2 billion impressions that expose users to ads that hide their political nature and avoid disclosing who paid for the ad.

7.7 Discussion

Across the ads where Facebook has to make an enforcement decision, we observe 61% more undetected ads (across political pages; Section 7.6.3) than detected ads (across all

pages; Section 7.5.1) within 14 days after their first activity. In addition, we observe that 55% of detections in the U.S. are likely false (Section 7.5.2). Translated into classification metrics, we estimate a precision of 0.45, a recall of 0.22, and subsequently an F_1 score of 0.29, all indicative of insufficiently accurate classification, and calling into question whether Facebook's enforcement is truly effective.

Incidentally, these error estimates are conservative and biased favorably towards better performance by Facebook. We quantify false negatives only across clearly political pages, where all ads must be declared, and enforcement is likely easier. If Facebook were to implement detection of every ad from such pages, our conservative estimate of the false negative rate would become zero. However, our estimate does not include potentially missed ads from other (non-political) advertisers; if these publish political ads without disclosing them, the ads would likely be more difficult to detect, given that they must be evaluated individually. Sosnovik and Goga [456] found 4% of 'strong political' ads to be undeclared, similar to our false negative rate. However, 7% of 'political' and 26% of 'marginally political' were also undeclared, and such ads were more often placed by NGOs, advocacy groups and charity organizations. This implies that we would also find a non-zero false negative rate if we were to extend our estimate to include individual ads from non-political actors. Nevertheless, our results present a baseline for ads currently missed by Facebook. Conversely, we conservatively estimate worldwide false positives based on our findings among advertisers in the United States. We found that the false negative rate was lowest in the United States; if a similar trend holds for false positives, Facebook's worldwide false positive rate is likely higher than in the U.S. Therefore, Facebook's global performance is likely worse than our estimate.

We now discuss five factors that enable more effective enforcement, and highlight how our findings suggest that Facebook's implementation is lacking in these areas. We also outline recommendations to Facebook for improving its enforcement and reduce erroneously missed or detected ads, as well as improve researchers' ability to audit its enforcement.

First, in terms of technical capability, Facebook's enforcement **approach** appears insufficient for the task of classifying political ads. Its automated moderation systems apparently do not learn or incorporate obvious signals of political intent, such as a page's self-categorization (Section 7.6.3), and this despite a high false positive rate (Section 7.5.2). Even if Facebook's review were more performant, the scale of its ad business means that low error rates still result in large absolute counts of missed political ads. *Recommendation:* Facebook should expand its enforcement approach to take the advertiser into account, e.g., by monitoring pages in political categories more strictly [157, 252]. Such simple, clearly enforceable rules could complement the current automated review.

Second, policy enforcement should be timely and come with appropriate **consequences** to prevent future violations. However, pages were still able to repeatedly run undeclared ads, even during the pause on political ads in the U.S. (Section 7.6.1). Moreover, violating ads sometimes run for a long period or are already inactive by the time of detection, resulting in large exposure (billions of impressions) before they are caught, if ever (Section 7.5.1). *Recommendation:* Facebook should ensure stricter consequences for

repeatedly violating advertisers, such as (temporarily) restricting them from running ads.

Third, enforcement must be **consistent** in order to be fair and effective for all users and advertisers. However, next to the overall enforcement errors that suggest inadequate reviewing resources, it appears that missed ads are more common outside the United States, where Facebook's enforcement suffers from higher false negative rates (Section 7.6.3), leaving users there more vulnerable to obscured political ads. *Recommendation:* Facebook should ensure consistent performance globally, independent of an ad's language. To capture cultural differences, they should engage with local governments, regulators and organizations to adapt policies and enforcement strategies to the local context [252, 451]. This includes identifying country-specific sensitive topics. Furthermore, they should mandate ad declaration worldwide, to ensure that no users in any country are unnecessarily left vulnerable to malicious political advertisers [365, 451].

Fourth, enforcement errors could result from insufficient ad **policies**. We find many largely non-political advertisers who appear to (possibly unintentionally) violate these policies and have detected ads (Section 7.6.2), even though they might have good intentions and be unaware that their ad was 'political.' This may be due to ambiguity in ad policies, in particular whether social issue ads "seek to influence public opinion." Our 'expert' annotators did not always agree on whether an ad was political (Section 7.5.2), suggesting that advertisers may also find this difficult, in particular as policies are spread out across many resources [3, 9, 10, 16, 23–25, 64, 125, 129, 197, 232, 233]. *Recommendation:* Facebook should further clarify and simplify its political ad policies, making it very obvious whether an ad is in scope or not [92, 451]. In addition, policies should be collected in one easily discoverable location [274, 333], with updates being clearly indicated and previous versions remaining available [92].

Finally, the quality of enforcement also affects the **transparency** into the political ad ecosystem that the Ad Library is meant to provide. Missed political ads disappear from the Ad Library once they become inactive, and additional metadata such as its spend and impressions are unavailable. Conversely, falsely detected non-political ads result in unrelated advertisers and ads appearing in the Ad Library, which may result in overestimating political ads on Facebook, and increases the (infrastructural and human) resources required to retrieve, process and analyze data for all advertisers (Section 7.4.5). *Recommendation:* Although we commend Facebook for their current transparency efforts, as they enable our audit and allow us to suggest improvements, they should expand transparency by including all ads in their archive and API to enable reproducible and scalable analysis of their enforcement [5, 120, 156, 164, 235, 250, 252, 253, 295, 404, 497]. For detected ads, they should also disclose which policy was violated and how they determined this [92], instead of the current binary signal.

However, changes to enforcement and transparency should be balanced with legitimate commercial and privacy concerns around sharing ad metadata, as well as consider adversarial counteractions from advertisers, who for example could attempt to evade efforts to identify them as a political actor (e.g., by selecting an unrelated page category). Actors beyond Facebook may therefore also need to intervene: legislators could

harmonize definitions of both political and issue ads across platforms [274] as well as set enforcement and transparency requirements [5, 120] that would be overseen by regulators [497] (Section 7.D). Ultimately, such regulatory pressure would entail a shift away from the current self-regulatory model to co-regulation [164, 497]: being allowed to self-regulate policies requires being able to enforce them well, which we show Facebook currently fails to achieve.

7.8 Conclusion

Through a large-scale collection of all ads from 215,030 pages with political ads over seven months, we conduct an audit of Facebook's political ad policy enforcement. We study whether this enforcement prevents negligent or malicious advertisers from weakening the integrity of the online political ad ecosystem by running political ads without disclosing them as required, and whether enforcement unnecessarily harms legitimate advertisers. Unfortunately, we find that Facebook's detection of political ads is flawed: Facebook misses more ads than they detect, and over half of those detected ads are incorrectly flagged. This enables advertisers to violate policies for an extended time or even evade bans on political ads. We attribute these flaws to limitations in Facebook's approach that does not sufficiently take into account the political intent of advertisers, allows pages to continue running violating ads, does not appear to be localized well in many countries, and is based on ambiguous policies that are harder to comply with and to enforce. These flaws then result in worse transparency into the online political ad ecosystem, as undisclosed and undetected political ads are neither accounted for in the summary statistics of the Ad Library Report, nor archived in the Ad Library so that the ads could be scrutinized after they become inactive. Yet, despite its flaws, it is also due to this transparency that we can audit Facebook's enforcement and formulate our recommendations to improve it: By being able to hold platforms accountable, we can work towards more secure online political speech.

7.A Ad Library web portal

We describe the metadata that is available through the Ad Library web portal, and annotate how the Ad Library web portal displays a declared and a detected ad (Figure 7.8).[9]

Within the Ad Library, each ad has a unique *archive ID*. For any ad, metadata consists of:

- its current *status* (active/inactive),
- the *start* and *end date* when an ad had active impressions,
- one or more of Facebook's advertising *platform*s where the ad is shown: currently Facebook, Instagram, Messenger, WhatsApp, and/or Audience Network,

[9]This interface differs from how a user sees an ad in their News Feed.

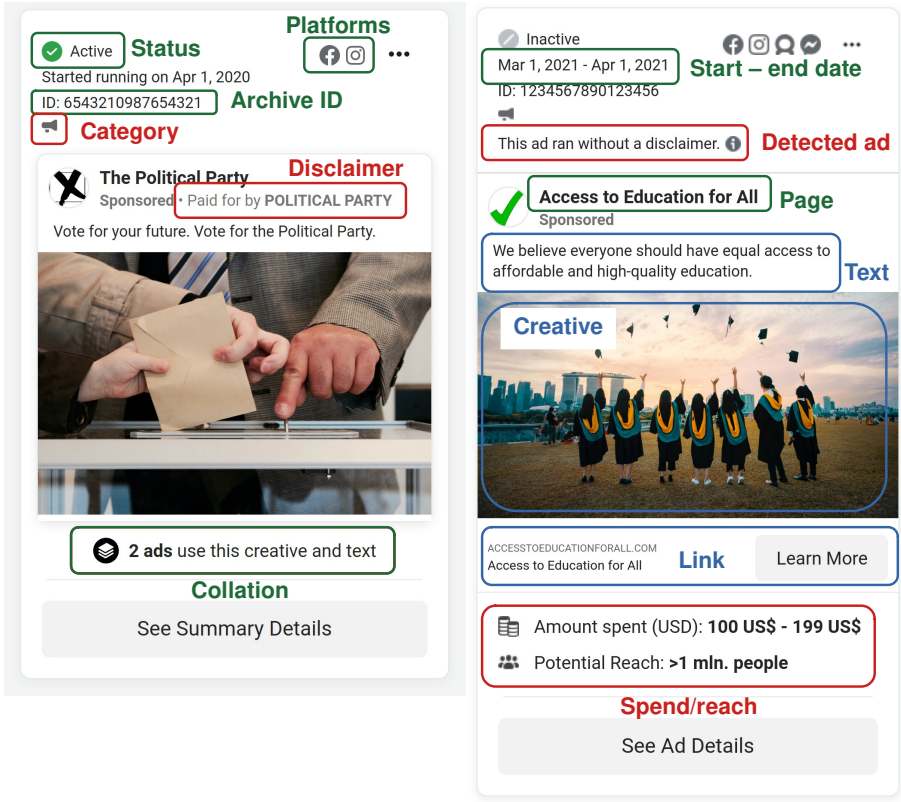


Figure 7.8: Annotated examples of a declared and a detected ad as displayed on the Ad Library web portal.

- the Facebook *page* that is running the ad,
- whether the ad contains a *reshared* post or *branded content*,
- whether the ad is *restricted* (e.g., by age) or *removed*.

The ad's contents consist of one or more *texts*, *creatives* (image or video), and *links*. Facebook groups ads with the same texts and creatives but different settings (such as active status, start/end dates, targeting) into one *collation*.

An ad may belong to a 'special ad *category*' [114]: social issues, elections or politics (worldwide) or housing, credit, or employment (U.S. and Canada). For ads about social issues, elections or politics, additional metadata consists of:

- for declared ads, the *disclaimer* used ("Paid for by"); or for detected ads, a message that "this ad ran without a disclaimer",
- binned estimates of *spend*, *reach*, and *impressions*, as well as distributions of impressions over age, gender, and region.

A Facebook page's metadata consists of (self-assigned) categories, its age, its like count, whether it is deleted, and whether it is verified. For a page that has completed the authorization process, additional metadata is available, which varies by country and advertiser type but may include the advertiser's name, address and website, as well as how Facebook verified the advertiser's identity.

7.B Data collection timeline

Figure 7.9 shows how our data collection and set evolved over time. We highlight events that affect either availability of political ads on Facebook, or of our data collection pipeline.

7.C Discarded pages

As mentioned in Section 7.4.2, the Ad Library Report contains a few highly active yet mostly non-political advertisers. These advertisers consume disproportionate scraping resources and skew our counts of observed ads. To reduce these two issues, we selected the largest global advertisers, i.e., those that had over 500 daily ads on average, and discarded those that we considered unlikely to intentionally publish political ads. Two annotators first categorized the pages that met our threshold (270 pages as of our last evaluation on December 7, 2020), and deliberated on which categories to exclude from scraping. We retained a page if we considered it possible for the page to still publish a genuine political ad.

Overall, we discarded 151 large pages (55.9%) (Table 7.6), mostly pages promoting a commercial product or service; an app, game or website; or clickbait content. Only 0.011% of their observed ads were political, supporting our decision to discard these pages. We observed nearly 5.9 million unique ads for these pages, but as we discarded pages during our measurement, this count is truncated. If we assume a constant daily ad rate, these pages would have accounted for nearly 15.4 million ads (35.6% of all ads in that case), yet only contributed 1,472 political ads (0.035% of all political ads), heavily skewing our data set. For these discarded pages, we assume that any political ad was accidental, and should be considered outside the scope of the political ad policy, i.e., a false positive.

We retained 119 large pages (44.1%). Only pages that were clearly related to political content (candidates, parties, etc.) had a meaningful share of political ads (72.0%). We opted to retain news/media pages that appear to not fall under Facebook's policy of exempting news organizations from declaring political ads [16, 24], although they published few political ads (0.35% of their ads). We also retained pages in categories that might (sometimes) be considered about 'social issues,' but their share of political ads was also low (below 0.3%). While we could have opted to also discard these latter two categories of pages, we assume that Facebook's enforcement systems might consider (some of) their ads to be within the scope of the political ad policy.

Table 7.6: Manual categorization of the largest advertisers in scope, grouped by our decision to disable further scraping or not. We also list the number of affected ads per group.

Page discarded (very unlikely political)

	of pages		of ads
Topic	# ▽	%	% pol.
Commercial product/service	63	23.3	0.007
Apps/games/websites	35	13.0	0.023
Clickbait	31	11.5	0.010
Housing/real estate	10	3.7	0.003
Employment	9	3.3	0.016
Other	3	1.1	0.038
All discarded pages	151	55.9	0.011
Total observed ads			5,862,808 ads
of which political			670 ads
Total estimated ads			15,438,700 ads
of which political			1,472 ads

Page retained (possibly political)

	of pages		of ads
Topic	# ▽	%	% pol.
News/media	69	25.6	0.353
Political content	20	7.4	72.0
Health	10	3.7	0.058
Solar panels	8	3.0	0.038
Government programs	5	1.9	0.271
Social issues	4	1.5	0.058
Other	3	1.1	0.020
All retained pages	119	44.1	5.57
Total observed ads			8,390,267 ads
of which political			467,130 ads

7.D Legal framework for online political and issue advertising

In major jurisdictions, election-related regulations have not been adapted yet to online political advertising specifically [252, 295]. In the United States, at the federal level reporting and funding disclosure is only required for ads that call to vote for a candidate or that are published by a candidate, PAC or party; moreover, these responsibilities lie with the advertiser rather than the platform [182]. The proposed *Honest Ads Act* would extend these requirements to all online political and 'issue' ads, where issue ads "relat[e] to a national legislative issue of public importance" [109], and mandate that platforms archive such ads [295]. The proposed *Social Media DATA Act* would mandate platforms to provide academic researchers with a library of ads and their targeting data [476]. In the European Union, major platforms signed the 'self-regulatory' *Code of Practice on Disinformation*, where they pledged to a.o. disclose political and issue ads (according to their own definitions), and publish them in ad archives [164, 227, 295, 497]. The *Digital Services Act* and the *European Democracy Action Plan* are slated to introduce similar but legally binding provisions and harmonize them across EU member states, some of which previously had no such requirements [147, 227, 252, 456]. More extensive requirements already exist in Brazil, where only candidates and parties are allowed to buy political ads and platforms must clearly indicate that an ad is political [439], and in Canada, where platforms are required to maintain a registry of political ads during elections [227].

7.E Topic codebook

For the manual analysis of false positives (Section 7.5.2), annotators developed the following codebook for describing the ad topic.

- *Related to politics and elections*
  - About a political figure/organization
  - About elections
  - By a political figure/organization
  - Insignificant reference to politics
- *Related to social issues*
  - Civil rights
  - Crime
  - Economy
  - Education
  - Environment
  - Food assistance
  - Government programs
  - Guns
  - Health

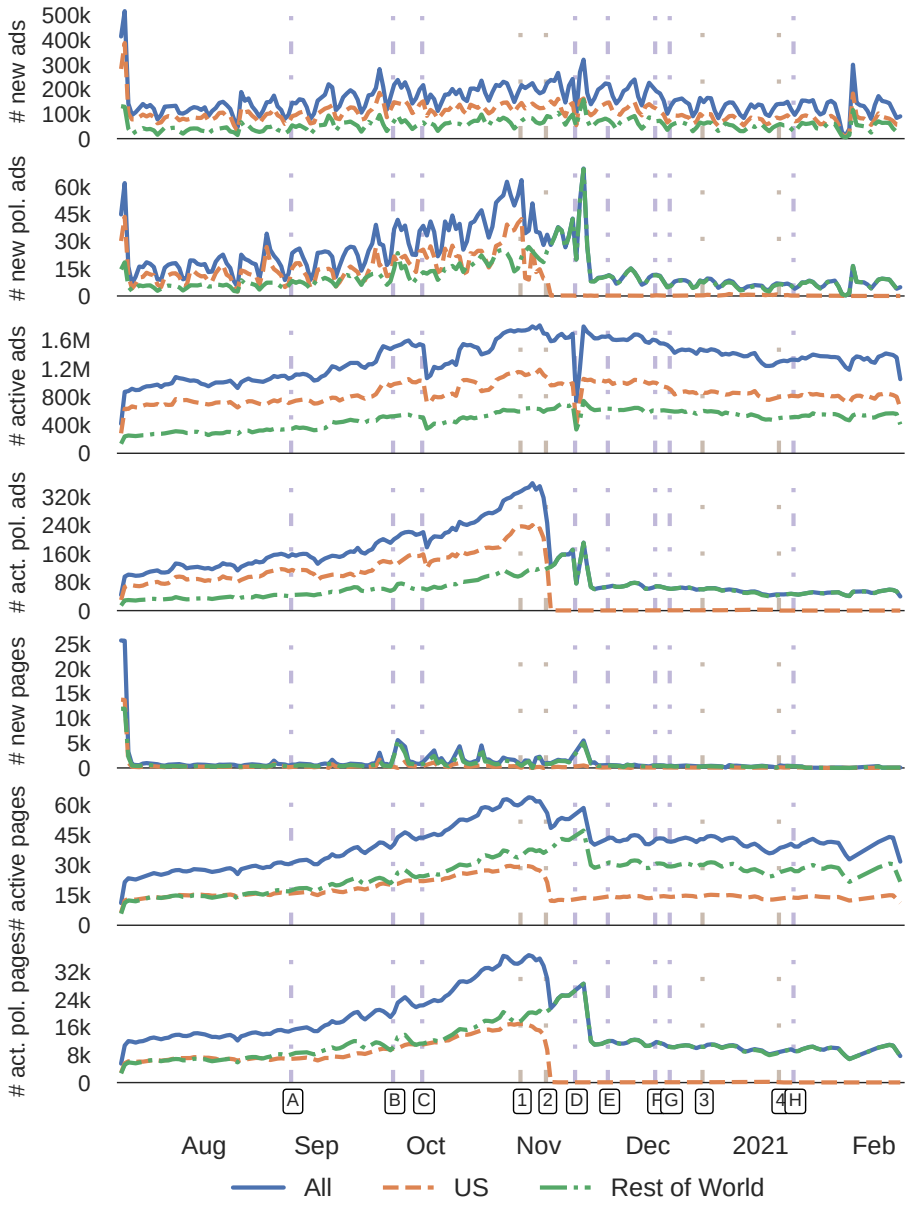
  - Immigration
  - Political Values and Governance
  - Security and Foreign Policy
- *Other*
  - Apps/games/websites
  - Arts/Personalities
  - COVID-19-related
  - Clickbait
  - Commercial product/service
  - Employment
  - Housing
  - News/media
  - Religion
  - Scams

7.F Page categories

Table 7.7 lists the page categories considered for each page class used in the classification for Section 7.6.2.

Table 7.7: Facebook page categories per page class.

Category name	ID
Political	
Political Candidate	842783295865930
Politician	1700
Political Party	2618
Political Organization	373543049350668
Government Official	1701
Government	
Public & Government Service	147714868971098
Government Organization	161422927240513
Public Service	139386576124160
Government Building	1032965636792826
City Hall	436168419731123
City	2404
Issue	
Nonprofit Organization	2603
Charity Organization	226326230802065
Labor Union	192775991124365
Environmental Conservation Organization	191523214199822
Non-Governmental Organization (NGO)	2235
Cause	2606
Media	
Media/News Company	2233
News & Media Website	2709
Newspaper	108366235907857
Broadcasting & Media Production Company	169056916473899
Magazine	1307
Publisher	191684877517919
TV Channel	1404
Media	1314020451960517
Books & Magazines	979978068761972



Facebook events

1	Oct 27, 2020	No new political ads accepted in U.S.
2	Nov 04, 2020	No political ads running in U.S.
3	Dec 16, 2020	Political ads running in Georgia
4	Jan 06, 2021	No political ads running in Georgia

Data collection events

A	Aug 25, 2020	Added 1 country
B	Sep 22, 2020	Added 33 countries
C	Sep 30, 2020	Discarded 96 pages
D	Nov 11, 2020	Crawler failure
E	Nov 20, 2020	Discarded 16 pages
F	Dec 03, 2020	Discarded 10 pages
G	Dec 07, 2020	Discarded 29 pages
H	Jan 10, 2021	No new pages added

Figure 7.9: Timeline of ad/page counts, with relevant events.

7.G Ad Library report dates

Table 7.8 lists the countries where Ad Library reports were available from the start of our data collection, while Table 7.9 lists the countries where reports were made available after the start of our data collection. We list the dates when Facebook first published reports, and when we first scraped ads for pages in these reports in our measurement. Finally, we mark countries where ads about social issues had to be declared during our measurement (column ⊙), and where Ad Library API data was available to us.

Table 7.8: Countries with Ad Library reports included from the start of our data collection.

Country (code)		Report start	Scrape start	⊙	API
Argentina	AR	2019-09-26	2020-07-09		
Austria	AT	2019-04-15	2020-07-09	✓	✓
Belgium	BE	2019-04-15	2020-07-09	✓	✓
Bulgaria	BG	2019-04-15	2020-07-09	✓	
Canada	CA	2019-06-25	2020-07-09	✓	✓
Croatia	HR	2019-04-15	2020-07-09	✓	
Cyprus	CY	2019-04-15	2020-07-09	✓	
Czechia	CZ	2019-04-15	2020-07-09	✓	
Denmark	DK	2019-04-15	2020-07-09	✓	
Estonia	EE	2019-04-15	2020-07-09	✓	
Finland	FI	2019-04-15	2020-07-09	✓	✓
France	FR	2019-04-15	2020-07-09	✓	✓
Germany	DE	2019-04-15	2020-07-09	✓	✓
Greece	GR	2019-04-15	2020-07-09	✓	
Hungary	HU	2019-04-15	2020-07-09	✓	
India	IN	2019-02-21	2020-07-09		✓
Ireland	IE	2019-04-15	2020-07-09	✓	✓
Israel	IL	2019-08-01	2020-07-09		
Italy	IT	2019-04-15	2020-07-09	✓	
Latvia	LV	2019-04-15	2020-07-09	✓	
Lithuania	LT	2019-04-15	2020-07-09	✓	
Luxembourg	LU	2019-04-15	2020-07-09	✓	✓
Malta	MT	2019-04-15	2020-07-09	✓	
Netherlands	NL	2019-04-15	2020-07-09	✓	✓
Poland	PL	2019-04-15	2020-07-09	✓	✓
Portugal	PT	2019-04-15	2020-07-09	✓	
Romania	RO	2019-04-15	2020-07-09	✓	
Singapore	SG	2019-09-26	2020-07-09	✓	
Slovakia	SK	2019-04-15	2020-07-09	✓	
Slovenia	SI	2019-04-15	2020-07-09	✓	
Spain	ES	2019-04-15	2020-07-09	✓	
Sri Lanka	LK	2020-05-18	2020-07-09		
Sweden	SE	2019-04-15	2020-07-09	✓	
Taiwan	TW	2019-11-11	2020-07-09	✓	
Ukraine	UA	2019-06-25	2020-07-09		✓
United Kingdom	GB	2018-11-29	2020-07-09	✓	✓
United States	US	2018-05-07	2020-07-09	✓	✓

Table 7.9: Countries with Ad Library reports included after the start of our data collection.

Country (code)		Report start	Scrape start	⊙	API
Australia	AU	2020-08-04	2020-09-22		
Belize	BZ	2020-08-04	2020-09-22		
Bolivia	BO	2020-08-04	2020-09-22		✓
Brazil	BR	2020-08-04	2020-09-22		✓
Burkina Faso	BF	2020-08-04	2020-09-22		
Chile	CL	2020-08-04	2020-09-22		
Colombia	CO	2020-08-04	2020-09-22		
Dominican Rep.	DO	2020-08-04	2020-09-22		
Ecuador	EC	2020-08-04	2020-09-22		
Georgia	GE	2020-08-04	2020-09-22		
Ghana	GH	2020-08-04	2020-09-22		
Guyana	GY	2020-08-04	2020-09-22		
Iceland	IS	2020-08-04	2020-09-22		
Indonesia	ID	2020-08-04	2020-09-22		
Ivory Coast	CI	2020-08-04	2020-09-22		
Japan	JP	2020-08-04	2020-09-22		
Kyrgyzstan	KG	2020-08-04	2020-09-22		
Malaysia	MY	2020-08-04	2020-09-22		
Mali	ML	2020-08-04	2020-09-22		
Mexico	MX	2020-08-04	2020-09-22		
Moldova	MD	2020-08-04	2020-09-22		
Mongolia	MN	2020-08-04	2020-09-22		
Montenegro	ME	2020-08-04	2020-09-22		
Myanmar	MM	2020-08-04	2020-09-22	✓	✓
New Zealand	NZ	2020-07-13	2020-08-25	✓	
North Macedonia	MK	2020-08-04	2020-09-22		
Palau	PW	2020-08-04	2020-09-22		
Philippines	PH	2020-08-04	2020-09-22		
Saint Vincent	VC	2020-08-04	2020-09-22		
Serbia	RS	2020-08-04	2020-09-22		
Seychelles	SC	2020-08-04	2020-09-22		
Suriname	SR	2020-08-04	2020-09-22		
Tanzania	TZ	2020-08-04	2020-09-22		
Turkey	TR	2020-08-04	2020-09-22		

Part III

Conclusion

8 Conclusion

In this dissertation, we took a critical view on common research practices within web security, using a meta-research standpoint to analyze how methods and data sets affect our ability to conduct valid and sound research into major security issues and ecosystems. We conclude with a reflection on enablers and challenges for research in our field, including future topics, that ultimately allow us to develop improved security solutions to better protect end users.

8.1 The importance of data sets

One underlying theme of the work presented in this dissertation is the importance of suitable and available data sets to conducting valid and groundbreaking research. Our *Tranco* work showed that researchers relied on domain rankings that exhibited undesirable properties for research, threatening the validity of the studies that depended on them. These rankings were also lacking on scientific tenets such as transparency and reproducibility, due to their opaque methods and lack of archives. With our contribution of the Tranco ranking, we have already made a step forward in the direction of improving upon these properties and providing a viable alternative to the research community.

Our work on automated decision-making systems highlighted the need for reliable data set access to be able to develop, evaluate, and audit such systems, and the need for adapting (research) methods to account for missing or unavailable data. In our *Avalanche* work, we considered the impact of data set unavailability, both at the level of individual domains and the domain set as a whole. This unavailability reflects the real-world constraint that data may be difficult to acquire and is likely to be incomplete. To account for individual domains with missing data, we found that an ensemble model trained on the different combinations of available data sets can robustly generate a prediction for every domain. Across the whole domain set, we found that data sets are partially interchangeable without significant performance loss, due to all of them capturing time-based patterns. However,

our solution does work better with more data being available, and seeking out additional data sets therefore remains a valuable effort. For our *Facebook* work, the data available in Facebook's main transparency tool (the Ad Library API) was insufficient for our purpose, as we needed to measure how long it took for Facebook to enforce their policy as well as discover those ads that Facebook failed to detect. We therefore developed a custom data collection pipeline using an alternative data source (the web portal), requiring additional engineering effort and continuous monitoring, in particular to allow for undocumented changes to that data source. The availability of this data was crucial for our findings on the speed and coverage of enforcement.

In terms of data access, the two topics that this dissertation covers are moving in opposite directions. Research into domain rankings is increasing and maturing, with both new and open ranking approaches being proposed [58, 352, 523] and further critical evaluation of existing approaches being performed [415]. In general, initiatives to make data on web security and privacy more open are gaining momentum, with the HTTP Archive and its yearly Web Almanac [511], a report on the current state of the web with chapters on security and privacy, being one prime example. This positive trend should be applauded and encouraged.

For both case studies on automated decision-making systems, the continued availability of crucial data is less certain. For takedown analyses, the compatibility of detailed WHOIS data with privacy regulation such as the GDPR remains difficult, and little progress appears to have been made in implementing alternatives that would guarantee continued availability for legitimate purposes [2, 517]. Solutions for DNS-based data remain mostly in the commercial space, although projects such as OpenINTEL [489] and SIE Europe [4] provide open access to vetted partners. For social media and advertising analyses, Facebook has shown itself to be increasingly resistant to providing researchers with access to data. Multiple researchers and organizations already identified technical [92, 118, 149, 157, 168] and policy-related [164, 478, 497] shortcomings in Facebook's current (political) ad transparency efforts. Crowdsourcing is an alternative way of conducting audit studies, and while it is more difficult to be representative or comprehensive, crowdsourcing has the benefit of ensuring that data relates to real users. However, Facebook has taken active steps to block browser extensions specifically designed for crowdsourcing ad targeting data from volunteers [231, 264, 325]. This includes our collaborators at New York University, whose accounts were disabled by Facebook over their Ad Observer browser extension [134, 504]. Facebook received backlash after this action, including from the Electronic Frontier Foundation, Mozilla, and the Federal Trade Commission [165, 285, 335]. Facebook will reportedly also shut down its CrowdTangle tool [31, 449], which provides research access to the public content of the most popular pages on its platform, and which has been used to study issues such as misinformation [159]. Other platforms are also reducing data access: there are concerns that Twitter will remove free access to its API for academic researchers [293], and Reddit recently disabled API access for the Pushshift service [301], which has been used extensively in academic research [79]. Combined, these actions present a significant challenge to studying harmful content on large online platforms, auditing their algorithms for bias, or auditing their policy enforcement. In general, the community should strive to reverse the negative trend

on data access for these audits, and may be helped by, a.o., legislative efforts to enforce increased transparency into automated systems that should also help to evaluate and improve their security-related properties.

8.2 Outlook and future work

Our work has highlighted the value of and need for reproducibility, transparency, and availability of data sets and research methods in general, to ultimately guarantee the validity and trustworthiness of the research that relies on them. We now give directions for future research that enhances the robustness of these desirable properties, and therefore ultimately of the research itself.

8.2.1 Domain rankings

With Tranco, we currently already provide a robust, well used, and increasingly important service to the community. However, new tools would make Tranco even more useful, and would guarantee its long-term availability and impact.

The space of domain rankings needs to be continuously monitored, to track both the arrival of new lists and the discontinuation of existing lists. Ideally, the new lists would cover multiple vantage points and traffic sources, e.g., being more indicative of web browsing instead of DNS queries. Concretely, the Chrome User Experience Report [116] and Cloudflare Radar [320] rankings are primary candidates for integration, but Tranco's list generation algorithm will need to adapted to account for their bucketed ranks (i.e., domains have a rank range, not an exact rank). We have ongoing work designing a domain ranking using passive DNS data in a privacy-preserving manner. This ranking would improve upon existing rankings in terms of transparency, as the ranking method would be fully transparent; and availability, as it would depend on a raw data source that is unlikely to disappear soon. In addition, new filters could prove useful for further aligning with researchers' needs. The filter on reachable domains, mentioned in the original Tranco paper, which would be based on responsiveness, status code, and content length, has not yet been implemented. This filter could further help to select domains representative of real websites. However, this would require a regular crawl of all domains in the ranking, a resource-intensive and time-sensitive task.

The Tranco website could be extended with a dashboard through which researchers can visually explore the Tranco ranking to better understand its properties, such as the long-term stability, relative importance of the component rankings, or composition in terms of, e.g., TLDs or organizations. To increase the resilience of Tranco, ensure that the previously generated lists remain retrievable and therefore ensure that the research that used those lists remains maximally reproducible, the lists could be duplicated on

a long-term stable (research) repository such as OSF[1] or the Internet Archive. With these features gradually being added to Tranco, the service can remain available, useful, relevant, and impactful for the foreseeable future.

An open question is how to cope with the differing vantage points and usages of existing and new rankings. Researchers should both be aware of how the purpose of their study interacts with the (desired) construction of a ranking, and be able to actually construct such a ranking that conforms to their needs. The current Tranco list incorporates rankings across a variety of vantage points, sourced from both web-based measurements and DNS traffic. Moreover, recent new ranking designs (e.g., SecRank [523]) rely on passive DNS data. This means that they might have bias towards Internet infrastructure domains, which may be popular as in regularly queried and accessed, but not as in regularly viewed through a web browser. While some types of studies are well served with such a ranking that includes infrastructure domains, some studies specifically study the web as viewed by users and therefore do not need or want infrastructure domains among their domain sample. When generating a customized Tranco list, we already provide the ability to select only web-focused rankings as well as filter on the Chrome User Experience Report to favor websites. This idea could be extended to generating separate rankings for web and infrastructural resources, based on a service classification that would indicate whether a domain rather hosts a website or is used as part of broader Internet infrastructure. This would provide researchers the ability to select the ranking that is most appropriate for their research. Recent work has also shown that popular websites differ significantly between countries and languages [32, 414]. We already provide filters on TLD or country data from the Chrome User Experience Report, but could search further data sources that represent country-specific popularity.

Both the new ranking proposals and existing rankings can benefit from further analysis and (long-term) evaluation to best understand their properties, biases, and suitability for research. One interesting avenue to further evaluate the accuracy and representativeness of these rankings could be to conduct a user study, where users worldwide indicate whether they know a particular domain that is ranked highly, and whether they actually consider it popular. This would complement research that measures domain popularity directly from user traffic. More boldly, we can ask ourselves whether a daily varying list is truly necessary, at least for research purposes. Perhaps, the community would be better served with a list that remains quasi unchanged over time, or at least only at a slow rate. We too preferred stability in the design of our Tranco list, averaging over 30 days to smooth out the high daily variability of some component lists. Similarly, an exact assignment of a rank to every domain might be unnecessary, and runs the risk that researchers assign too much importance to the minute difference in ranks between certain domains, as they may not represent a large difference in traffic. Ruth et al. [415] also found that papers mostly ignore the ranks and just require a representative sample of domains. The bucketing approach that the Chrome User Experience Report already takes might be a good middle ground, as a notion of relative popularity remains while not claiming a large precision for the ranks. Working towards a ranking that best meets

[1] https://osf.io/

the requirements of researchers in our field would best be a community effort to properly understand their needs.

8.2.2 Large-scale web measurements

More broadly, in the space of large-scale web measurements, there is a good momentum for further improving upon these properties, and there would be an equal benefit from improved data sets and tools, but there remain several unsolved or unexplored challenges.

We already analyzed the properties of third-party domain classification services [486], which are used to contextualize measurements by website category, but no open, research-oriented solution is available to this date. Given the large number of domains to be classified, most current services rely on an automated classification using, e.g., text-based features. Maintaining such a service may be challenging and resource-intensive at scale. An alternative approach may be to use the human-contributed mappings of domains to their operating entities in wiki-based services such as DBpedia [296] (based on Wikipedia) and Wikidata [500], to some extent returning to the roots of domain classification in human-maintained directories such as DMOZ and OpenDNS [486]. These (machine-readable) data sets contain a 'website' field for each entity or article, alongside attributes that include that entity's name and a category. These data sets are curated by the Wikipedia/Wikidata contributors, suggesting they may be of higher quality than automated solutions, though this remains to be evaluated. The data is likely to be much richer in terms of describing the entity that operates a domain. However, coverage may be limited and skewed to entities of encyclopedic notability, or simply interests of contributors [217]. These data sources may be lacking data on infrastructure domains in particular, as these are typically not the 'official' website of a company or entity and therefore unlikely to be described on Wikipedia or Wikidata. Nevertheless, for those domains where an article is available, they immediately are linked to the full structured metadata available on Wikipedia and Wikidata, and even beyond through linked unique identifiers in other data sources. Ultimately, this links domains to the Semantic Web [86], readily enriching them with other data sources. A service like Tranco could even be expanded to provide domain intelligence in a semantic format to link back to these data sources. Such a service could then be used to gain further insights into domains through the metadata available for them, or select domain samples in a more advanced way, e.g., choosing websites of companies listed on a given stock exchange. This service could even integrate with security services such as threat intelligence platforms, and help investigators to better contextualize domains.

Thinking big, one could imagine a "toolbox" where researchers can find resources for measurements for which there is (a degree of) agreement within the community that they represent current best practices to achieve maximally valid and reproducible research, yielding a form of standardization across research studies. Such a toolbox can stretch from suggesting appropriate data sets like Tranco or the classification service discussed above, all the way to providing crawling infrastructure that is preconfigured with verified parameters for maximally representative data collection. This selection of parameters

would control for variables such as the geographical location of a crawl, type of network (residential, cloud, university, …), user agent, crawl duration, authenticated crawls, stateful or stateless crawls, page interaction, cloaking circumvention, internal pages, and so on. Inspiration can come from the Internet measurement community, where collaborative platforms for data collection of network data are already common [73], e.g., M-Lab [200], RIPE Atlas [407], or CAIDA Ark [110].

This toolbox would be contributed to by multiple types of research. Studies thoroughly analyzing the properties, biases, and impact of data sets, crawler setups, and configuration, for example by executing multiple crawls where these parameters are varied in a controlled environment, would serve as a basis for the best practices proposed to researchers, contributing through a meta-research perspective. Studies developing new data sets and crawlers, or extending and improving existing solutions, would serve as the concrete tools provided to researchers. The insights and tools obtained through these experiments would provide further insights into the practices that enable sound research. Moreover, they would strongly encourage reproducibility, as the selected parameters and crawling results could be easily shared, similar to how the configuration of a customized Tranco ranking is available on that ranking's page. Artifact evaluation committees could more easily reproduce results, and other researchers could more easily leverage other researchers' results. Repeating a study could be as simple as rerunning the data collection using the same infrastructure and parameters. Ultimately, this could become a large-scale collaborative crawling platform, where researchers share data sets, results, and best practices, for the benefit of the community.

8.2.3 Automated decision-making systems

Given the fundamental questions that the usage of automated decision-making systems brings about, in terms of these systems' accuracy, efficacy, and ethical behavior, also in security-related applications, further research that critically assesses these systems for their appropriateness remains crucial. Essential to this task is transparency into these systems and sufficient access to the data that serves as their inputs and outputs [501]. We join the calls for increased transparency into these systems, in areas such as online political advertising [5, 120, 156, 164, 235, 250, 252, 253, 295, 404, 497], misinformation [314, 330] and algorithmic fairness [306, 396, 508]. This transparency would enable further research into the way automated decision-making systems make their decisions, the accuracy of these systems' results, and ultimately the impact they have. This also spurs the development of new research techniques to address these questions. However, if platforms do not readily provide transparency, there is also still a need for new research methods that allow researchers to collect data on these systems autonomously, such as the ad scraper that we developed for our audit of political ads on Facebook. The data sets that these data collection pipelines produce should also be available to other researchers to allow for reproducibility as well as building upon prior work and gaining novel insights.

There are legitimate concerns attached to sharing data publicly. User privacy must be preserved, and 'over-sharing' data might cause personally identifiable information or other

sensitive information to be leaked. For example, too fine-grained targeting data may reveal a person's location or interests, or reveal other sensitive attributes, e.g., their political beliefs or sexual orientation if these are attributes that can be targeted or inferred from targeting data. This is a delicate challenge, and prior work has demonstrated the feasibility of deanonymization attacks on large-scale data set releases [356]. Proposed transparency standards, such as the Universal Ad Transparency standard [155], already incorporate these privacy concerns, e.g., by not revealing parameters if they target very small audiences. Moreover, there is a concern that more transparency into security systems can give attackers more direct insight in how to circumvent them [52, 278]. Overall, good practices should be learned from existing transparency requirements in other domains, such as data sharing in the context of clinical trials or environmental impacts, which has already enabled researchers to impact policymaking [361]. However, transparency should not be restricted to academic researchers, but must be more widely offered to industry practitioners, independent researchers, journalists, non-profit organizations, and the public at large. This would maximally enable critical investigations into how these systems work, and the development of new improvements to these systems.

Making data sets more widely available can have auxiliary benefits for exploring other security issues and malicious ecosystems as well, allowing defenders to develop better solutions based on more transparent data [344]. As one concrete example, advocacy about full transparency for all ads published on large platforms has focused on the benefits for increased accountability in the realm of political advertising [5, 120, 156, 235, 250, 252, 253, 295, 404, 497]. However, broader ad (meta)data would also be very useful for cybercrime studies into abuse related to online advertising, such as our ongoing study on deceptive affiliate marketing. Having such ad data would allow us to discover the advertisements that redirect to deceptive affiliate marketing products and services, analyze how quickly they are taken down, find major intermediaries in the hosting chain between advertisement and product landing page, and understand the dynamics of ad spend and targeting for these advertisements. After this improved understanding of the ecosystem, we can go on to developing defenses that protect users from seeing these ads and helping platforms in removing these ads. As shown previously, automated models benefit from more and better data, and the large ad data sets would be useful as training data for models detecting these deceptive ads. While researchers already use self-collected ad data for analyzing malicious or deceptive advertising [483, 530, 531], having an ad archive would guarantee the completeness, comprehensiveness, and representativeness of the ad collection.

While data transparency is not the only or a full solution to avoiding harms in automated decision-making systems, as this still requires curating the data sets and models to account for or remove any harmful biases, transparency is a crucial first step towards enabling us to understand which harms exist and why they might exist. Transparency enables us to study how platforms develop and deploy their automated decision-making systems, and how other parties such as advertisers use and potentially abuse these systems, therefore allowing us to hold all of these parties accountable. Such accountability then creates positive incentives: for platforms to improve their systems to, e.g., reduce biases or improve policy enforcement, and for the other parties to reduce harmful practices such

as selecting discriminatory targeting profiles or not complying with policies. Improved enforcement and compliance then lead to our final goal and the intent for studying these automated decision-making systems: achieving better security of these systems and the ecosystems they support, and ensuring that their beneficial potential starts outweighing the risks and vulnerabilities that we know to exist when they are deployed in practice.

8.3 Closing thoughts

Through our work in this dissertation, we contributed to analyzing and improving the data sets and methods available to our research community. Having access to soundly constructed, transparently developed, and better understood data sets and methods helps to increase the validity of research. Ultimately, this can make (web) security research a more scientifically grounded practice, by providing better resources to conduct robust web measurements and analyses, therefore avoiding any biases introduced by flaws in the data sets or methods. The open availability of data sets and methods also allows for better reproducibility of research results, to confirm whether their observations (continue to) genuinely represent the state of web security, therefore enabling more rigorous science that is valid, trustworthy and can be relied upon. Next to these more abstract improvements, we hope that with wider availability, these data sets and methods will enable more inclusive research, through more comprehensive and representative studies of web security issues, with better coverage of, a.o., more countries, demographics, and audiences in general, that improve the opportunities to observe the various web security issues that affect web users globally. These yield us more complete and thorough insights into malicious online practices that ultimately allow us to develop better security solutions that also more fairly protect users across all populations, and helping to make the web a more secure place for all.

Bibliography

[1] *'Avalanche' network dismantled in international cyber operation*. Press release. Europol, Dec. 1, 2016. URL: https://www.europol.europa.eu/newsroom/news/%E2%80%98avalanche%E2%80%99-network-dismantled-in-international-cyber-operation.

[2] *2021-06-21 - EPDP Phase 1 Policy Implementation (follow-up on previous GAC consensus advice)*. ICANN Governmental Advisory Committee. June 21, 2021. URL: https://gac.icann.org/advice/itemized/2021-06-21-epdp-phase-1-policy-implementation-follow-up-on-previous-gac-consensus-advice.

[3] *5 Things to Remember About Political and Issue Advertising Around the US 2020 Election*. Facebook for Business. Oct. 26, 2020. URL: https://www.facebook.com/business/news/facebook-ads-restriction-2020-us-election.

[4] *A Breakthrough European Data Sharing Collective to Fight Cybercrime*. SIE Europe UG. 2021. URL: https://www.sie-europe.net/.

[5] *A comprehensive plan to innovate democracy in Europe: Civil society vision for the European Democracy Action Plan*. European Partnership for Democracy, Sept. 2020. URL: https://epd.eu/wp-content/uploads/2020/09/a-civil-society-vision-for-the-european-democracy-action-plan-input-paper.pdf.

[6] *A Look at Facebook and US 2020 Elections*. Report. Facebook, Dec. 2020. URL: https://about.fb.com/wp-content/uploads/2020/12/US-2020-Elections-Report.pdf.

[7] G. Aaron and R. Rasmussen. *Global Phishing Survey: Trends and Domain Name Use in 2H2009*. APWG Industry Advisory. Anti-Phishing Working Group, May 2010. URL: https://docs.apwg.org/reports/APWG_GlobalPhishingSurvey_2H2009.pdf.

[8] J. Abbink and C. Doerr. "Popularity-based Detection of Domain Generation Algorithms". In: *12th International Conference on Availability, Reliability and Security*. ARES '17. 2017, 79. DOI: 10.1145/3098954.3107008.

[9] *About Ads About Social Issues, Elections or Politics*. Facebook Business Help Center. Mar. 3, 2021. URL: https://www.facebook.com/business/help/167836590566506.

[10] *About Social Issues*. Facebook Business Help Center. Apr. 7, 2021. URL: https://www.facebook.com/business/help/214754279118974.

[11] *About the Ad Library*. Facebook Business Help Center. Sept. 2, 2020. URL: https://www.facebook.com/business/help/2405092116183307.

[12] S. Abt and H. Baier. "Are We Missing Labels? A Study of the Availability of Ground-Truth in Network Security Research". In: *3rd International Workshop on Building Analysis Datasets and Gathering Experience Returns for Security*. BADGERS '14. 2014. DOI: 10.1109/badgers.2014.11.

[13] M. Abu Rajab, F. Monrose, A. Terzis, and N. Provos. "Peeking through the Cloud: DNS-Based Estimation and Its Applications". In: *6th International Conference on Applied Cryptography and Network Security*. ACNS '08. 2008, pp. 21–38.

[14] abuse.ch. *SinkDB*. 2019. URL: https://sinkdb.abuse.ch/.

[15] G. Acar, C. Eubank, S. Englehardt, M. Juarez, A. Narayanan, and C. Diaz. "The Web Never Forgets: Persistent Tracking Mechanisms in the Wild". In: *21st ACM SIGSAC Conference on Computer and Communications Security*. CCS '14. 2014, pp. 674–689.

[16] *Ad Authorization Exemptions and How They Work*. Facebook Business Help Center. Dec. 15, 2020. URL: https://www.facebook.com/business/help/387111852028957.

[17] *Ad Library*. Facebook. 2021. URL: https://www.facebook.com/ads/library/.

[18] *Ad Library API*. Facebook. 2021. URL: https://www.facebook.com/ads/library/api/.

[19] *Ad Library Report*. Facebook. 2021. URL: https://www.facebook.com/ads/library/report/.

[20] A. Adadi and M. Berrada. "Peeking Inside the Black-Box: A Survey on Explainable Artificial Intelligence (XAI)". In: *IEEE Access* 6 (2018), pp. 52138–52160. DOI: 10.1109/ACCESS.2018.2870052.

[21] L. A. Adamic and B. A. Huberman. "Zipf's law and the Internet". In: *Glottometrics* 3 (2002), pp. 143–150.

[22] E. Adie and W. Roe. "Altmetric: enriching scholarly content with article-level discussion and metrics". In: *Learned Publishing* 26.1 (2013), pp. 11–17. DOI: 10.1087/20130103.

[23] *Ads About Social Issues, Elections or Politics*. Facebook Business Help Center. Mar. 10, 2021. URL: https://www.facebook.com/business/help/1838453822893854.

[24] *Advertising Policies: Restricted Content: Ads About Social Issues, Elections or Politics*. Facebook. 2021. URL: https://www.facebook.com/policies/ads/restricted_content/political.

[25] *Advertising Policies: Restricted Content: Disclaimers for Ads About Social Issues, Elections or Politics*. Facebook. 2021. URL: https://www.facebook.com/policies/ads/restricted_content/disclaimers.

[26] M. Aertsen, M. Korczyński, G. C. M. Moura, S. Tajalizadehkhoob, and J. van den Berg. "No Domain Left behind: Is Let's Encrypt Democratizing Encryption?" In: *2017 Applied Networking Research Workshop*. ANRW '17. 2017, pp. 48–54. DOI: 10.1145/3106328.3106338.

[27] S. S. Ahmad, M. D. Dar, M. F. Zaffar, N. Vallina-Rodriguez, and R. Nithyanand. "Apophanies or Epiphanies? How Crawlers Impact Our Understanding of the Web". In: *The Web Conference 2020*. WWW '20. 2020, pp. 271–280. DOI: 10.1145/3366423.3380113.

[28] A. Akella and N. Taft. *IMC 2014 Decision on Public Reviews*. 2014. URL: https://conferences.sigcomm.org/imc/2014/news3.html.

[29] D. Akhawe, A. Barth, P. E. Lam, J. Mitchell, and D. Song. "Towards a Formal Foundation of Web Security". In: *23rd IEEE Computer Security Foundations Symposium*. CSF '10. 2010, pp. 290–304. DOI: 10.1109/CSF.2010.27.

[30] E. S. Alashwali, P. Szalachowski, and A. Martin. "Exploring HTTPS security inconsistencies: A cross-regional perspective". In: *Computers & Security* 97 (2020), p. 101975. DOI: 10.1016/j.cose.2020.101975.

[31] D. Alba. "Meta Pulls Support for Tool Used to Keep Misinformation in Check". In: *Bloomberg* (June 23, 2022). URL: https://www.bloomberg.com/news/articles/2022-06-23/meta-pulls-support-for-tool-used-to-keep-misinformation-in-check.

[32] T. Alby and R. Jäschke. "Analyzing the Web: Are Top Websites Lists a Good Choice for Research?" In: *26th International Conference on Theory and Practice of Digital Libraries*. TPDL '22. 2022, pp. 11–25. DOI: 10.1007/978-3-031-16802-4_2.

[33] Alexa Internet, Inc. *Does Alexa have a list of its top-ranked websites?* Archived on April 4, 2016. Jan. 15, 2016. URL: https://web.archive.org/web/20160404003433/https://support.alexa.com/hc/en-us/articles/200449834-Does-Alexa-have-a-list-of-its-top-ranked-websites- (visited on 03/02/2018).

[34] Alexa Internet, Inc. *Does Alexa have a list of its top-ranked websites?* Archived on March 11, 2017. Jan. 26, 2017. URL: https://web.archive.org/web/20170311160137/https://support.alexa.com/hc/en-us/articles/200449834-Does-Alexa-have-a-list-of-its-top-ranked-websites- (visited on 04/24/2018).

[35] Alexa Internet, Inc. *Global Top Sites*. Dec. 16, 2008. URL: https://web.archive.org/web/20081216072512/http://www.alexa.com:80/site/ds/top_sites.

[36] Alexa Internet, Inc. *How are Alexa's traffic rankings determined?* Nov. 9, 2017. URL: https://support.alexa.com/hc/en-us/articles/200449744.

[37] Alexa Internet, Inc. *What exactly is the Alexa Traffic Panel?* May 1, 2018. URL: https://support.alexa.com/hc/en-us/articles/200080859.

[38] Alexa Support (@AlexaSupport). *Thanks to customer feedback, the top 1M sites is temporarily available again. We'll provide notice before updating the file in the future*. Nov. 22, 2016. URL: https://twitter.com/Alexa_Support/status/801167423726489600.

[39] Alexa Support (@AlexaSupport). *Yes, the top 1m sites file has been retired.* Nov. 21, 2016. URL: https://twitter.com/Alexa_Support/status/800755671784308736.

[40] *Alexa Traffic Rank - Chrome Web Store*. Nov. 2018. URL: https://chrome.google.com/webstore/detail/alexa-traffic-rank/cknebhggccemgcnbidipinkifmmegdel.

[41] M. Ali. "Measuring and Mitigating Bias and Harm in Personalized Advertising". In: *15th ACM Conference on Recommender Systems*. RecSys '21. 2021, pp. 869–872. DOI: 10.1145/3460231.3473895.

[42] M. Ali, P. Sapieżyński, M. Bogen, A. Korolova, A. Mislove, and A. Rieke. "Discrimination through Optimization: How Facebook's Ad Delivery Can Lead to Biased Outcomes". In: *Proceedings of the ACM on Human-Computer Interaction* 3.CSCW (Nov. 2019). DOI: 10.1145/3359301.

[43] M. Ali, P. Sapieżyński, A. Korolova, A. Mislove, and A. Rieke. "Ad Delivery Algorithms: The Hidden Arbiters of Political Messaging". In: *14th ACM International Conference on Web Search and Data Mining*. WSDM '21. 2021, pp. 13–21. DOI: 10.1145/3437963.3441801.

[44] M. Allman and V. Paxson. "Issues and Etiquette Concerning Use of Shared Measurement Data". In: *7th ACM SIGCOMM Conference on Internet Measurement*. IMC '07. 2007, pp. 135–140. DOI: 10.1145/1298306.1298327.

[45] A. Almomani, B. B. Gupta, S. Atawneh, A. Meulenberg, and E. Almomani. "A Survey of Phishing Email Filtering Techniques". In: *IEEE Communications Surveys & Tutorials* 15.4 (2013), pp. 2070–2090. DOI: 10.1109/SURV.2013.030713.00020.

[46] E. Alowaisheq, P. Wang, S. Alrwais, X. Liao, X. Wang, T. Alowaisheq, X. Mi, S. Tang, and B. Liu. "Cracking the Wall of Confinement: Understanding and Analyzing Malicious Domain Take-downs". In: *26th Annual Network and Distributed System Security Symposium*. NDSS '19. 2019. DOI: 10.14722/ndss.2019.23243.

[47] S. Alrwais, X. Liao, X. Mi, P. Wang, X. Wang, F. Qian, R. Beyah, and D. McCoy. "Under the Shadow of Sunshine: Understanding and Detecting Bulletproof Hosting on Legitimate Service Provider Networks". In: *2017 IEEE Symposium on Security and Privacy*. SP '17. 2017, pp. 805–823. DOI: 10.1109/SP.2017.32.

[48] Amazon Web Services, Inc. *Alexa Top Sites*. Mar. 15, 2018. URL: https://aws.amazon.com/alexa-top-sites/.

[49] Amazon Web Services, Inc. *Alexa Web Information Service*. Aug. 3, 2018. URL: https://aws.amazon.com/awis/.

[50] Amazon Web Services, Inc. *AWS IP Address Ranges*. Apr. 25, 2018. URL: https://docs.aws.amazon.com/general/latest/gr/aws-ip-ranges.html/.

[51] *Analyzing PyPI package downloads*. Python Packaging User Guide. 2022. URL: https://packaging.python.org/en/latest/guides/analyzing-pypi-package-downloads/.

[52] M. Ananny and K. Crawford. "Seeing without knowing: Limitations of the transparency ideal and its application to algorithmic accountability". In: *New Media & Society* 20.3 (2018), pp. 973–989. DOI: 10.1177/1461444816676645.

[53] A. Andreou, M. Silva, F. Benevenuto, O. Goga, P. Loiseau, and A. Mislove. "Measuring the Facebook Advertising Ecosystem". In: *26th Annual Network and Distributed System Security Symposium*. NDSS '19. 2019. DOI: 10.14722/ndss.2019.23280.

[54] M. Antonakakis, R. Perdisci, D. Dagon, W. Lee, and N. Feamster. "Building a Dynamic Reputation System for DNS". In: *19th USENIX Security Symposium*. USENIX Security '10. 2010, pp. 273–289.

[55] M. Antonakakis, R. Perdisci, Y. Nadji, N. Vasiloglou, S. Abu-Nimeh, W. Lee, and D. Dagon. "From Throw-Away Traffic to Bots: Detecting the Rise of DGA-Based Malware". In: *21st USENIX Security Symposium*. USENIX Security '12. 2012, pp. 491–506.

[56] M. Antonakakis et al. "Understanding the Mirai botnet". In: *26th USENIX Security Symposium*. USENIX Security '17. 2017, pp. 1093–1110.

[57] S. Aonzo, Y. Han, A. Mantovani, and D. Balzarotti. "32nd USENIX Security Symposium". In: *USENIX Security '23*. 2023.

[58] W. Aqeel, B. Chandrasekaran, A. Feldmann, and B. M. Maggs. "On Landing and Internal Web Pages: The Strange Case of Jekyll and Hyde in Web Performance Measurement". In: *20th Internet Measurement Conference*. IMC '20. 2020, pp. 680–695. DOI: 10.1145/3419394.3423626.

[59] D. Arp, E. Quiring, F. Pendlebury, A. Warnecke, F. Pierazzi, C. Wressnegger, L. Cavallaro, and K. Rieck. "Dos and Don'ts of Machine Learning in Computer Security". In: *31st USENIX Security Symposium*. USENIX Security '22. 2022, pp. 3971–3988.

[60] M. De-Arteaga, R. Fogliato, and A. Chouldechova. "A Case for Humans-in-the-Loop: Decisions in the Presence of Erroneous Algorithmic Scores". In: *2020 CHI Conference on Human Factors in Computing Systems*. CHI '20. 2020, pp. 1–12. DOI: 10.1145/3313831.3376638.

[61] *Artifact Review and Badging*. Version 1.1. Association for Computing Machinery. Aug. 24, 2020. URL: https://www.acm.org/publications/policies/artifact-review-and-badging-current.

[62] H. Asghari, M. Ciere, and M. J. van Eeten. "Post-Mortem of a Zombie: Conficker Cleanup After Six Years". In: *24th USENIX Security Symposium*. USENIX Security '15. 2015, pp. 1–16.

[63] *Automated Data Collection Terms*. Facebook. Apr. 15, 2010. URL: https://www.facebook.com/apps/site_scraping_tos_terms.php.

[64] *Availability for Ads About Social Issues, Elections or Politics*. Facebook Business Help Center. Mar. 11, 2021. URL: https://www.facebook.com/business/help/2150157295276323.

[65] *Avalanche 1,2,3...* The Shadowserver Foundation. Dec. 2, 2018. URL: http://blog.shadowserver.org/news/avalanche-123/.

[66] *Avalanche Stats by Subregion*. The Shadowserver Foundation. URL: https://avalanche.shadowserver.org/stats/ (visited on 03/08/2019).

[67] *Avalanche year two, this time with Andromeda*. The Shadowserver Foundation. Dec. 4, 2017. URL: http://blog.shadowserver.org/news/avalanche-year-two-this-time-with-andromeda/.

[68] Y. Aydin. *Étude nationale portant sur la sécurité de l'espace numérique français 2017*. French. National study. Available in English at https://www.economie.gouv.fr/files/2017_National_Study_Cybersecurity.pdf. Ministères économiques et financiers, Oct. 4, 2017, pp. 4–14. URL: https://www.economie.gouv.fr/files/2017_Etude_nationale_securite_numerique.pdf.

[69] M. Bailey, D. Dittrich, E. Kenneally, and D. Maughan. "The Menlo Report". In: *IEEE Security & Privacy Magazine* 10.2 (Mar. 2012), pp. 71–75. DOI: 10.1109/msp.2012.52.

[70] V. Bajpai, O. Bonaventure, K. Claffy, and D. Karrenberg. "Encouraging Reproducibility in Scientific Research of the Internet (Dagstuhl Seminar 18412)". In: *Dagstuhl Reports* 8.10 (2019), pp. 41–62. DOI: 10.4230/DagRep.8.10.41.

[71] V. Bajpai, A. Brunstrom, A. Feldmann, W. Kellerer, A. Pras, H. Schulzrinne, G. Smaragdakis, M. Wählisch, and K. Wehrle. "The Dagstuhl Beginners Guide to Reproducibility for Experimental Networking Research". In: *ACM SIGCOMM Computer Communication Review* 49.1 (Feb. 2019), pp. 24–30. DOI: 10.1145/3314212.3314217.

[72] V. Bajpai, M. Kühlewind, J. Ott, J. Schönwälder, A. Sperotto, and B. Trammell. "Challenges with Reproducibility". In: *Reproducibility Workshop*. Reproducibility '17. 2017, pp. 1–4. DOI: 10.1145/3097766.3097767.

[73] V. Bajpai and J. Schönwälder. "A Survey on Internet Performance Measurement Platforms and Related Standardization Efforts". In: *IEEE Communications Surveys & Tutorials* 17.3 (2015), pp. 1313–1341. DOI: 10.1109/COMST.2015.2418435.

[74] D. Balenson, T. Benzel, E. Eide, D. Emmerich, D. Johnson, J. Mirkovic, and L. Tinnel. "Toward Findable, Accessible, Interoperable, and Reusable Cybersecurity Artifacts". In: *15th Workshop on Cyber Security Experimentation and Test*. CSET '22. 2022, pp. 65–70. DOI: 10.1145/3546096.3546104.

[75] D. Balzarotti. *System Security Circus*. 2022. URL: https://www.s3.eurecom.fr/~balzarot/security-circus/.

[76] J. Bandy. "Problematic Machine Behavior: A Systematic Literature Review of Algorithm Audits". In: *Proceedings of the ACM on Human-Computer Interaction* 5.CSCW1 (Apr. 2021). DOI: 10.1145/3449148.

[77] E. Bansal. *Intel Owl Release v1.0.0*. The Honeynet Project. July 5, 2020. URL: https://www.honeynet.org/2020/07/05/intel-owl-release-v1-0-0/.

[78] T. Barabosch, A. Wichmann, F. Leder, and E. Gerhards-Padilla. "Automatic extraction of domain name generation algorithms from current malware". In: *IST-111/RSY-026 Symposium on Information Assurance and Cyber Defence*. NATO Science & Technology Organization, 2012. ISBN: 978-92-837-0177-4.

[79] J. Baumgartner, S. Zannettou, B. Keegan, M. Squire, and J. Blackburn. "The Pushshift Reddit Dataset". In: *Proceedings of the International AAAI Conference on Web and Social Media* 14.1 (May 2020), pp. 830–839. DOI: 10.1609/icwsm.v14i1.7347.

[80] *Beginner's Guide to Using NCCS Data*. Urban Institute, National Center for Charitable Statistics. Dec. 2018. URL: https://nccs.urban.org/sites/default/files/2018-12/Guide%20to%20Using%20NCCS%20Data.pdf.

[81] T. Benzel. "Security and Privacy Research Artifacts: Are We Making Progress?" In: *IEEE Security & Privacy* 21.01 (Jan. 2023), pp. 4–6. DOI: 10.1109/MSEC.2022.3222887.

[82] T. Benzel and F. Stajano. "IEEE Euro S&P: The Younger Sibling Across the Pond Following in Oakland's Footsteps". In: *IEEE Security & Privacy* 18.3 (2020), pp. 6–7. DOI: 10.1109/MSEC.2020.2980180.

[83] D. Beraldo, S. Milan, J. de Vos, C. Agosti, B. Nadalic Sotic, R. Vliegenthart, S. Kruikemeier, L. P. Otto, S. A. M. Vermeer, X. Chu, and F. Votta. "Political advertising exposed: tracking Facebook ads in the 2021 Dutch elections". In: *Internet Policy Review* (Mar. 11, 2021). URL: https://policyreview.info/articles/news/political-advertising-exposed-tracking-facebook-ads-2021-dutch-elections/1543.

[84] E. Berger. *CSRankings: Computer Science Rankings*. 2022. URL: https://csrankings.org/.

[85] E. Berger, S. M. Blackburn, C. Brodley, H. V. Jagadish, K. S. McKinley, M. A. Nascimento, M. Shin, K. Wang, and L. Xie. "GOTO Rankings Considered Helpful". In: *Communications of the ACM* 62.7 (June 2019), pp. 29–30. DOI: 10.1145/3332803.

[86] T. Berners-Lee, J. Hendler, and O. Lassila. "The Semantic Web". In: *Scientific American* 284.5 (2001), pp. 34–43.

[87] R. Beverly and M. Allman. "Findings and Implications from Data Mining the IMC Review Process". In: *ACM SIGCOMM Computer Communication Review* 43.1 (Jan. 2012), pp. 22–29. DOI: 10.1145/2427036.2427040.

[88] L. Bilge, E. Kirda, C. Kruegel, M. Balduzzi, and S. Antipolis. "EXPOSURE: Finding Malicious Domains Using Passive DNS Analysis". In: *18th Annual Network & Distributed System Security Symposium*. NDSS '11. 2011, pp. 1–17.

[89] L. Bilge, S. Sen, D. Balzarotti, E. Kirda, and C. Kruegel. "Exposure: A Passive DNS Analysis Service to Detect and Report Malicious Domains". In: *ACM Transactions on Information and System Security* 16.4 (Apr. 2014), pp. 1–28. DOI: 10.1145/2584679.

[90] C. Blundo, S. Cimato, and B. Masucci. "Secure Metering Schemes". In: *Network Security*. Ed. by S. C.-H. Huang, D. MacCallum, and D.-Z. Du. Springer US, 2010, pp. 1–32.

[91] boker et al. *Domain seized*. Dec. 27, 2018. URL: https://www.namepros.com/threads/domain-seized.1116091/.

[92] S. E. Bolden, B. McKernan, and J. Stromer-Galley. *Facebook Political Advertising Transparency Report*. Illuminating. Oct. 5, 2020. URL: https://news.illuminating.ischool.syr.edu/2020/10/06/facebook-political-advertising-transparency-report/.

[93] O. Bonaventure. "The January 2017 issue". In: *ACM SIGCOMM Computer Communication Review* 47.1 (Jan. 2017), pp. 1–3. DOI: 10.1145/3041027.3041028.

[94] K. Borgolte, C. Kruegel, and G. Vigna. "Meerkat: Detecting Website Defacements through Image-based Object Recognition". In: *24th USENIX Security Symposium*. USENIX Security '15. 2015, pp. 595–610.

[95] J. Bosso. *How does your location affect your online privacy?* Avast. Jan. 5, 2022. URL: https://blog.avast.com/location-and-online-privacy-avast.

[96] M. Botacin, F. Ceschin, R. Sun, D. Oliveira, and A. Grégio. "Challenges and pitfalls in malware research". In: *Computers & Security* 106, 102287 (July 2021). DOI: 10.1016/j.cose.2021.102287.

[97] N. Boucher and R. Anderson. "Talking Trojan: Analyzing an Industry-Wide Disclosure". In: *2022 ACM Workshop on Software Supply Chain Offensive Research and Ecosystem Defenses*. SCORED '22. 2022, pp. 83–92. DOI: 10.1145/3560835.3564555.

[98] *Brave Search passes 2.5 billion queries in its first year, and debuts Goggles feature that allows users to choose their own search rankings*. Brave. June 22, 2022. URL: https://brave.com/search-anniversary/.

[99] G. Brewer. *What is Tranco, Quantcast, Majestic and Umbrella Numbers?* BuiltWith. Aug. 10, 2018. URL: https://kb.builtwith.com/general-questions/what-is-tranco-quantcast-majestic-and-umbrella-numbers/.

[100] J. Brodkin. "How ISPs can sell your Web history — and how to stop them". In: *Ars Technica* (Mar. 24, 2017). URL: https://arstechnica.com/information-technology/2017/03/how-isps-can-sell-your-web-history-and-how-to-stop-them/.

[101] BuiltWith Pty Ltd. *Alexa Certified Site Metrics Usage Statistics*. Sept. 2018. URL: https://trends.builtwith.com/analytics/Alexa-Certified-Site-Metrics.

[102] *Call For Papers*. IEEE Symposium on Security and Privacy. 2017. URL: https://www.ieee-security.org/TC/SP2018/cfpapers.html.

[103] *Call For Papers*. IEEE Symposium on Security and Privacy. 2023. URL: https://sp2024.ieee-security.org/cfpapers.html.

[104] *Call For Posters, ACM IMC 2019, Amsterdam, The Netherlands*. Association for Computing Machinery. 2019. URL: https://conferences.sigcomm.org/imc/2019/call-for-posters/.

[105] Q. Cao, M. Sirivianos, X. Yang, and T. Pregueiro. "Aiding the Detection of Fake Accounts in Large Scale Social Online Services". In: *9th USENIX Conference on Networked Systems Design and Implementation*. NSDI '12. 2012, pp. 197–210.

[106] T. E. Carroll, D. Manz, T. Edgar, and F. L. Greitzer. "Realizing Scientific Methods for Cyber Security". In: *2012 Workshop on Learning from Authoritative Security Experiment Results*. LASER '12. 2012, pp. 19–24. DOI: 10.1145/2379616.2379619.

[107] D. Cassel, S.-C. Lin, A. Buraggina, W. Wang, A. Zhang, L. Bauer, H.-C. Hsiao, L. Jia, and T. Libert. "OmniCrawl: Comprehensive Measurement of Web Tracking With Real Desktop and Mobile Browsers". In: *Proceedings on Privacy Enhancing Technologies* 2022.1 (Sept. 2021), pp. 227–252. DOI: 10.2478/popets-2022-0012.

[108] L. Cavallaro, J. Kinder, F. Pendlebury, and F. Pierazzi. "Are Machine Learning Models for Malware Detection Ready for Prime Time?" In: *IEEE Security & Privacy* 21.2 (2023), pp. 53–56. DOI: 10.1109/MSEC.2023.3236543.

[109] G. Cecere, C. Jean, V. Lefrere, and C. E. Tucker. *Tradeoffs in Automated Political Advertising Regulation: Evidence from the COVID-19 Pandemic*. 2020. DOI: 10.2139/ssrn.3603341. SSRN: 3603341.

[110] Center for Applied Internet Data Analysis. *Archipelago (Ark) Measurement Infrastructure*. 2020. URL: https://www.caida.org/projects/ark/.

[111] O. Cetin, C. Gañán, L. Altena, T. Kasama, D. Inoue, K. Tamiya, Y. Tie, K. Yoshioka, and M. van Eeten. "Cleaning Up the Internet of Evil Things: Real-World Evidence on ISP and Consumer Efforts to Remove Mirai". In: *26th Annual Network and Distributed System Security Symposium*. NDSS '19. 2019. DOI: 10.14722/ndss.2019.23438.

[112] *CFP Changes for the 2024 Conference*. IEEE Symposium on Security and Privacy. 2023. URL: https://sp2024.ieee-security.org/changes-cfp.html.

[113] H. Chitkara. "To fix social media, senators turn to a research transparency bill". In: *Protocol* (Sept. 14, 2022). URL: https://www.protocol.com/bulletins/platform-accountability-act-senate.

[114] *Choosing a Special Ad Category*. Facebook Business Help Center. Feb. 19, 2021. URL: https://www.facebook.com/business/help/298000447747885.

[115] S. Christey and B. Martin. *Buying Into the Bias: Why Vulnerability Statistics Suck*. Black Hat USA 2013 Briefings, July 11, 2013.

[116] *Chrome User Experience Report*. Google Inc. Dec. 10, 2021. URL: https://developers.google.com/web/tools/chrome-user-experience-report.

[117] Y. T. Chua, S. Parkin, M. Edwards, D. Oliveira, S. Schiffner, G. Tyson, and A. Hutchings. "Identifying Unintended Harms of Cybersecurity Countermeasures". In: *2019 APWG Symposium on Electronic Crime Research*. eCrime '19. 2019.

[118] J. Chuang. *Data Collection Log – EU Ad Transparency Report*. Mozilla. 2019. URL: https://adtransparency.mozilla.org/eu/log/.

[119] Cisco Umbrella. *Umbrella Popularity List*. 2016. URL: https://s3-us-west-1.amazonaws.com/umbrella-static/index.html.

[120] *Civil society coalition calls for meaningful transparency on all ads*. European Partnership for Democracy, Sept. 8, 2020. URL: https://epd.eu/2020/09/08/pressreleaseadstransparency/.

[121] A. Clauset, C. R. Shalizi, and M. E. J. Newman. "Power-Law Distributions in Empirical Data". In: *SIAM Review* 51.4 (2009), pp. 661–703. DOI: 10.1137/070710111.

[122] R. Clayton and T. Mansfield. "A Study of Whois Privacy and Proxy Service Abuse". In: *13th Annual Workshop on the Economics of Information Security*. 2014.

[123] C. Collberg and T. A. Proebsting. "Repeatability in Computer Systems Research". In: *Communications of the ACM* 59.3 (Feb. 2016), pp. 62–69. DOI: 10.1145/2812803.

[124] Common Crawl Foundation. *Common Crawl*. URL: https://commoncrawl.org/.

[125] *Comply With Local Laws Governing Ads About Social Issues, Elections or Politics*. Facebook Business Help Center. Apr. 10, 2020. URL: https://www.facebook.com/business/help/534095800360404.

[126] *CORE Rankings Portal*. Computing Research and Education Association of Australasia. 2022. URL: https://www.core.edu.au/conference-portal.

[127] S. Costanza-Chock, I. D. Raji, and J. Buolamwini. "Who Audits the Auditors? Recommendations from a Field Scan of the Algorithmic Auditing Ecosystem". In: *5th ACM Conference on Fairness, Accountability, and Transparency*. FAccT '22. 2022, pp. 1571–1583. DOI: 10.1145/3531146.3533213.

[128] R. Costas, Z. Zahedi, and P. Wouters. "Do "altmetrics" correlate with citations? Extensive comparison of altmetric indicators with citations from a multidisciplinary perspective". In: *Journal of the Association for Information Science and Technology* 66.10 (2015), pp. 2003–2019. DOI: 10.1002/asi.23309.

[129] *Create Disclaimers and Link Ad Accounts*. Facebook Business Help Center. Jan. 5, 2021. URL: https://www.facebook.com/business/help/488070228549681.

[130] J. Crowcroft, S. Keshav, and N. McKeown. "Viewpoint Scaling the Academic Publication Process to Internet Scale". In: *Communications of the ACM* 52.1 (Jan. 2009), pp. 27–30. DOI: 10.1145/1435417.1435430.

[131] CrowdTangle Team. *CrowdTangle*. 2021. URL: https://www.crowdtangle.com/.

[132] A. Cuevas, F. Miedema, K. Soska, N. Christin, and R. van Wegberg. "Measurement by Proxy: On the Accuracy of Online Marketplace Measurements". In: *31st USENIX Security Symposium*. USENIX Security '22. 2022, pp. 2153–2170.

[133] R. R. Curtin, A. B. Gardner, S. Grzonkowski, A. Kleymenov, and A. Mosquera. "Detecting DGA Domains with Recurrent Neural Networks and Side Information". In: *14th International Conference on Availability, Reliability and Security*. ARES '19. 2019, 20. DOI: 10.1145/3339252.3339258.

[134] Cybersecurity for Democracy. *Ad Observer*. 2021. URL: https://adobserver.org/.

[135] L. Daigle. *WHOIS Protocol Specification*. RFC 3912. RFC Editor, Sept. 2004.

[136] S. Danziger, J. Levav, and L. Avnaim-Pesso. "Extraneous factors in judicial decisions". In: *Proceedings of the National Academy of Sciences* 108.17 (2011), pp. 6889–6892. DOI: 10.1073/pnas.1018033108.

[137] S. Das, J. Werner, M. Antonakakis, M. Polychronakis, and F. Monrose. "SoK: The Challenges, Pitfalls, and Perils of Using Hardware Performance Counters for Security". In: *2019 IEEE Symposium on Security and Privacy*. SP '19. 2019, pp. 20–38. DOI: 10.1109/SP.2019.00021.

[138] A. Datta, A. Datta, J. Makagon, D. K. Mulligan, and M. C. Tschantz. "Discrimination in Online Advertising: A Multidisciplinary Inquiry". In: *Proceedings of the 1st Conference on Fairness, Accountability and Transparency*. Vol. 81. Proceedings of Machine Learning Research. 2018, pp. 20–34. URL: https://proceedings.mlr.press/v81/datta18a.html.

[139] J. Davis. *Do top conferences contain well cited papers or junk?* 2019. DOI: 10.48550/arxiv.1911.09197. arXiv: 1911.09197.

[140] "Declaration of Special Agent Aaron O. Francis in support of application for an emergency temporary restraining order and order to show cause re preliminary injunction". In: *United States of America v. "flux" a/k/a "ffhost", and "flux2" a/k/a "ffhost2"*. District Court, Western District of Pennsylvania, Nov. 2016. URL: https://www.justice.gov/opa/page/file/915231/download.

[141] N. Demir, M. Große-Kampmann, T. Urban, C. Wressnegger, T. Holz, and N. Pohlmann. "Reproducibility and Replicability of Web Measurement Studies". In: *ACM Web Conference 2022*. WWW '22. 2022, pp. 533–544. DOI: 10.1145/3485447.3512214.

[142] DENIC. *DENIC Putting Extensive Changes into Force for .DE Whois Lookup Service by 25 May 2018*. May 2018. URL: https://www.denic.de/en/whats-new/press-releases/article/denic-putting-extensive-changes-into-force-for-de-whois-lookup-service-as-of-25-may-2018/.

[143] N. Diakopoulos. "Accountability in Algorithmic Decision Making". In: *Communications of the ACM* 59.2 (Jan. 2016), pp. 56–62. DOI: 10.1145/2844110.

[144] K. Dickersin. "The Existence of Publication Bias and Risk Factors for Its Occurrence". In: *JAMA* 263.10 (Mar. 1990), pp. 1385–1389. DOI: 10.1001/jama.1990.03440100097014.

[145] R. DiResta, L. Edelson, B. Nyhan, and E. Zuckerman. "It's Time to Open the Black Box of Social Media". In: *Scientific American* (Apr. 28, 2022). URL: https://www.scientificamerican.com/article/its-time-to-open-the-black-box-of-social-media/.

[146] D. Dittrich and E. Kenneally. *The Menlo Report: Ethical Principles Guiding Information and Communication Technology Research*. U.S. Department of Homeland Security, Aug. 2012.

[147] T. Dobber, S. Kruikemeier, E. P. Goodman, N. Helberger, and S. Minihold. *Effectiveness of Online Political Ad Disclosure Labels: Empirical Findings*. Institute for Information Law, University of Amsterdam; Rutgers Institute for Information Policy and Law, Mar. 8, 2021. URL: https://www.uva-icds.net/wp-content/uploads/2021/03/Summary-transparency-discloures-experiment_update.pdf.

[148] N. Doty. *Mitigating Browser Fingerprinting in Web Specifications*. W3C Editor's Draft. W3C, July 2018. URL: https://w3c.github.io/fingerprinting-guidance/.

[149] P. Duke. *What it's like to actually use Facebook's ad transparency tools*. Online Political Transparency Project. Oct. 30, 2020. URL: https://medium.com/online-political-transparency-project/7accf22f4ba7.

[150] Z. Durumeric, D. Adrian, A. Mirian, M. Bailey, and J. A. Halderman. "A Search Engine Backed by Internet-Wide Scanning". In: *22nd ACM SIGSAC Conference on Computer and Communications Security*. CCS '15. 2015, pp. 542–553. DOI: 10.1145/2810103.2813703.

[151] Z. Durumeric, E. Wustrow, and J. A. Halderman. "ZMap: Fast Internet-wide Scanning and Its Security Applications". In: *22nd USENIX Security Symposium*. USENIX Security '13. 2013, pp. 605–620.

[152] S. Eberz, K. B. Rasmussen, V. Lenders, and I. Martinovic. "Evaluating Behavioral Biometrics for Continuous Authentication: Challenges and Metrics". In: *2017 ACM on Asia Conference on Computer and Communications Security*. ASIA CCS '17. 2017, pp. 386–399. DOI: 10.1145/3052973.3053032.

[153] L. Edelson. *Audit of Facebook ad transparency finds missed political ads*. Online Political Transparency Project. Oct. 22, 2020. URL: https://medium.com/online-political-transparency-project/603f95027cc6.

[154] L. Edelson. *Publishing Facebook ad data (redux)*. Online Political Transparency Project. Nov. 13, 2020. URL: https://medium.com/online-political-transparency-project/ff071c41c12e.

[155] L. Edelson, J. Chuang, E. F. Fowler, M. M. Franz, and T. Ridout. *A Standard for Universal Digital Ad Transparency*. Knight First Amendment Institute, Dec. 9, 2021. URL: https://knightcolumbia.org/content/a-standard-for-universal-digital-ad-transparency.

[156] L. Edelson, E. F. Fowler, and J. Chuang. "We need universal digital ad transparency now". In: *TechCrunch* (Oct. 16, 2020). URL: https://techcrunch.com/2020/10/16/we-need-universal-digital-ad-transparency-now/.

[157] L. Edelson, T. Lauinger, and D. McCoy. "A Security Analysis of the Facebook Ad Library". In: *2020 IEEE Symposium on Security and Privacy*. SP '20. 2020, pp. 661–678. DOI: 10.1109/SP40000.2020.00084.

[158] L. Edelson and D. McCoy. "How Facebook Hinders Misinformation Research". In: *Scientific American* (Sept. 22, 2021). URL: https://www.scientificamerican.com/article/how-facebook-hinders-misinformation-research/.

[159] L. Edelson, M.-K. Nguyen, I. Goldstein, O. Goga, D. McCoy, and T. Lauinger. "Understanding Engagement with U.S. (Mis)Information News Sources on Facebook". In: *21st Internet Measurement Conference*. IMC '21. 2021, pp. 444–463. DOI: 10.1145/3487552.3487859.

[160] L. Edelson, S. Sakhuja, R. Dey, and D. McCoy. *An Analysis of United States Online Political Advertising Transparency*. 2019. arXiv: 1902.04385 [cs.SI].

[161] *Election Integrity at Facebook*. Facebook for Business. 2021. URL: https://www.facebook.com/business/m/election-integrity.

[162] G. Engelen, V. Rimmer, and W. Joosen. "Troubleshooting an Intrusion Detection Dataset: the CICIDS2017 Case Study". In: *2021 IEEE Security and Privacy Workshops*. SPW '21. 2021, pp. 7–12. DOI: 10.1109/SPW53761.2021.00009.

[163] S. Englehardt and A. Narayanan. "Online Tracking: A 1-million-site Measurement and Analysis". In: *23rd ACM SIGSAC Conference on Computer and Communications Security*. CCS '16. 2016, pp. 1388–1401. DOI: 10.1145/2976749.2978313.

[164] *ERGA Report on disinformation: Assessment of the implementation of the Code of Practice*. European Regulators Group for Audiovisual Media Services, May 2020. URL: https://erga-online.eu/wp-content/uploads/2020/05/ERGA-2019-report-published-2020-LQ.pdf.

[165] M. Erwin. *Why Facebook's claims about the Ad Observer are wrong*. Mozilla. Aug. 4, 2021. URL: https://blog.mozilla.org/en/mozilla/news/why-facebooks-claims-about-the-ad-observer-are-wrong/.

[166] J. Espinoza and C. Criddle. "Meta bosses look at political ads ban in Europe". In: *Financial Times* (Mar. 30, 2023). URL: https://www.ft.com/content/1a133b5c-35f9-4776-99fc-7c02095ff2aa.

[167] *Exempt Organizations Business Master File Extract*. Internal Revenue Service. 2021. URL: https://www.irs.gov/charities-non-profits/exempt-organizations-business-master-file-extract-eo-bmf.

[168] *Facebook Ads Library Assessment*. Office of the French Ambassador for Digital Affairs. 2020. URL: https://disinfo.quaidorsay.fr/en/facebook-ads-library-assessment.

[169] *Facebook Open Research & Transparency*. Meta. 2022. URL: https://fort.fb.com/.

[170] *Facebook response to the public consultation for the European Democracy Action Plan*. Facebook, Sept. 15, 2020. URL: https://about.fb.com/wp-content/uploads/sites/10/2020/09/Facebook_Response_European_Democracy_Action_Plan_2020.09.15.pdf.

[171] M. Faddoul, G. Chaslot, and H. Farid. *A Longitudinal Analysis of YouTube's Promotion of Conspiracy Videos*. 2020. DOI: 10.48550/arxiv.2003.03318. arXiv: 2003.03318.

[172] I. Faizullabhoy and A. Korolova. "Facebook's Advertising Platform: New Attack Vectors and the Need for Interventions". In: *2nd Workshop on Technology and Consumer Protection*. ConPro '18. 2018. URL: https://www.ieee-security.org/TC/SPW2018/ConPro/papers/faizullabhoy-conpro18.pdf.

[173] D. Fanelli. "Is science really facing a reproducibility crisis, and do we need it to?" In: *Proceedings of the National Academy of Sciences* 115.11 (2018), pp. 2628–2631. DOI: 10.1073/pnas.1708272114.

[174] Farsight Security. *DNSDB*. https://www.dnsdb.info/.

[175] Á. Feal, P. Vallina, J. Gamba, S. Pastrana, A. Nappa, O. Hohlfeld, N. Vallina-Rodriguez, and J. Tapiador. "Blocklist Babel: On the Transparency and Dynamics of Open Source Blocklisting". In: *IEEE Transactions on Network and Service Management* 18.2 (2021), pp. 1334–1349. DOI: 10.1109/TNSM.2021.3075552.

[176] D. G. Feitelson. *"We do not appreciate being experimented on": Developer and Researcher Views on the Ethics of Experiments on Open-Source Projects*. 2021. DOI: 10.48550/arxiv.2112.13217. arXiv: 2112.13217.

[177] A. Feldmann. "Experiences from the Sigcomm 2005 European Shadow PC Experiment". In: *ACM SIGCOMM Computer Communication Review* 35.3 (July 2005), pp. 97-7–102. DOI: 10.1145/1070873.1070889.

[178] M. Felegyhazi, C. Kreibich, and V. Paxson. "On the Potential of Proactive Domain Blacklisting". In: *3rd USENIX Conference on Large-scale Exploits and Emergent Threats: Botnets, Spyware, Worms, and More*. LEET '10. 2010.

[179] A. P. Felt, R. Barnes, A. King, C. Palmer, C. Bentzel, and P. Tabriz. "Measuring HTTPS Adoption on the Web". In: *26th USENIX Security Symposium*. USENIX Security '17. 2017, pp. 1323–1338.

[180] A. J. Ferrante. "The impact of GDPR on WHOIS: Implications for businesses facing cybercrime". In: *Cyber Security: A Peer-Reviewed Journal* 2.2 (2018), pp. 143–148. ISSN: 2398-5100.

[181] M. Flittner, M. N. Mahfoudi, D. Saucez, M. Wählisch, L. Iannone, V. Bajpai, and A. Afanasyev. "A Survey on Artifacts from CoNEXT, ICN, IMC, and SIGCOMM Conferences in 2017". In: *ACM SIGCOMM Computer Communication Review* 48.1 (Apr. 2018), pp. 75–80. DOI: 10.1145/3211852.3211864.

[182] E. F. Fowler, M. M. Franz, and T. N. Ridout. "Online Political Advertising in the United States". In: *Social Media and Democracy: The State of the Field, Prospects for Reform*. Ed. by N. Persily and J. A. Tucker. SSRC Anxieties of Democracy. Cambridge University Press, 2020, pp. 111–138. ISBN: 9781108890960. DOI: 10.1017/9781108890960.

[183] E. Frachtenberg. "Research artifacts and citations in computer systems papers". In: *PeerJ Computer Science* 8 (Feb. 2022), e887. DOI: 10.7717/peerj-cs.887.

[184] E. Frachtenberg and N. Koster. "A survey of accepted authors in computer systems conferences". In: *PeerJ Computer Science* 6 (Sept. 2020), e299. DOI: 10.7717/peerj-cs.299.

[185] J. Fraenkel and B. Grofman. "The Borda Count and its real-world alternatives: Comparing scoring rules in Nauru and Slovenia". In: *Australian Journal of Political Science* 49.2 (2014), pp. 186–205. DOI: 10.1080/10361146.2014.900530.

[186] Freenom. *Free and paid domains*. 2017. URL: https://www.freenom.com/en/freeandpaiddomains.html.

[187] B. Friedman and H. Nissenbaum. "Bias in Computer Systems". In: *ACM Transactions on Information Systems* 14.3 (July 1996), pp. 330–347. DOI: 10.1145/230538.230561.

[188] S. Frier. "Facebook's Political Rule Blocks Ads for Bush's Beans, Singers Named Clinton". In: *Bloomberg* (July 2, 2018). URL: https://www.bloomberg.com/news/articles/2018-07-02/facebook-s-algorithm-blocks-ads-for-bush-s-beans-singers-named-clinton.

[189] T. Frosch, M. Kührer, and T. Holz. "Predentifier: Detecting Botnet C&C Domains From Passive DNS Data". In: *Advances in IT Early Warning*. Ed. by M. Zeilinger, P. Schoo, and E. Hermann. Fraunhofer Verlag, Feb. 2013, pp. 78–90. ISBN: 978-3-8396-0474-8. URL: http://publica.fraunhofer.de/documents/N-227985.html.

[190] D. Fujs, A. Mihelič, and S. L. R. Vrhovec. "The Power of Interpretation: Qualitative Methods in Cybersecurity Research". In: *14th International Conference on Availability, Reliability and Security*. ARES '19. 2019. DOI: 10.1145/3339252.3341479.

[191] D. Gaffney and J. N. Matias. "Caveat emptor, computational social science: Large-scale missing data in a widely-published Reddit corpus". In: *PLOS ONE* 13.7 (July 2018). DOI: 10.1371/journal.pone.0200162.

[192] D. Gale. "Facebook's Problem With Veterans". In: *The Wall Street Journal* (Aug. 7, 2018). URL: https://www.wsj.com/articles/facebooks-problem-with-veterans-1533682511.

[193] S. Gallagher. "New "Quad9" DNS service blocks malicious domains for everyone". In: *Ars Technica* (Nov. 16, 2017). URL: https://arstechnica.com/information-technology/2017/11/new-quad9-dns-service-blocks-malicious-domains-for-everyone/.

[194] O. Gasser, Q. Scheitle, P. Foremski, Q. Lone, M. Korczyński, S. D. Strowes, L. Hendriks, and G. Carle. "Clusters in the Expanse: Understanding and Unbiasing IPv6 Hitlists". In: *2018 Internet Measurement Conference*. IMC '18. 2018, pp. 364–378. DOI: 10.1145/3278532.3278564.

[195] A. Gee-Clough. *Mirror, Mirror, on the Wall, Who's the Fairest (website) of Them all?* DomainTools. Apr. 28, 2022. URL: https://www.domaintools.com/resources/blog/mirror-mirror-on-the-wall-whos-the-fairest-website-of-them-all.

[196] M. Gelbmann. *Extension of our website sample set*. W3Techs. Mar. 31, 2020. URL: https://w3techs.com/blog/entry/extension_of_our_website_sample_set.

[197] *Get Authorized to Run Ads About Social Issues, Elections or Politics*. Facebook Business Help Center. Mar. 3, 2021. URL: https://www.facebook.com/business/help/208949576550051.

[198] A. Ghosh, G. Venkatadri, and A. Mislove. "Analyzing Political Advertisers' Use of Facebook's Targeting Features". In: *3rd Workshop on Technology and Consumer Protection*. ConPro '19. 2019. URL: https://www.ieee-security.org/TC/SPW2019/ConPro/papers/ghosh-conpro19.pdf.

[199] D. Gidwani. *Herding Cattle: ThreatConnect's Vision for Better Intel Feeds*. ThreatConnect. Feb. 19, 2020. URL: https://threatconnect.com/blog/cal-2-4-introducing-cal-feeds/.

[200] P. Gill, C. Diot, L. Y. Ohlsen, M. Mathis, and S. Soltesz. "M-Lab: User Initiated Internet Data for the Research Community". In: *ACM SIGCOMM Computer Communication Review* 52.1 (Mar. 2022), pp. 34–37. DOI: 10.1145/3523230.3523236.

[201] E. Glazer and J. Horwitz. "Facebook Curbs Incentives to Sell Political Ads Ahead of 2020 Election". In: *The Wall Street Journal* (May 23, 2019). URL: https://www.wsj.com/articles/facebook-ends-commissions-for-political-ad-sales-11558603803.

[202] D. Gomes, J. Miranda, and M. Costa. "A Survey on Web Archiving Initiatives". In: *International Conference on Theory and Practice of Digital Libraries*. TPDL '11. 2011, pp. 408–420. DOI: 10.1007/978-3-642-24469-8_41.

[203] J. González Cabañas, Á. Cuevas, A. Arrate, and R. Cuevas. "Does Facebook Use Sensitive Data for Advertising Purposes?" In: *Communications of the ACM* 64.1 (Dec. 2020), pp. 62–69. DOI: 10.1145/3426361.

[204] J. González Cabañas, Á. Cuevas, and R. Cuevas. "FDVT: Data Valuation Tool for Facebook Users". In: *2017 CHI Conference on Human Factors in Computing Systems*. CHI '17. 2017, pp. 3799–3809. DOI: 10.1145/3025453.3025903.

[205] J. González Cabañas, Á. Cuevas, and R. Cuevas. "Unveiling and Quantifying Facebook Exploitation of Sensitive Personal Data for Advertising Purposes". In: *27th USENIX Security Symposium*. USENIX Security '18. 2018, pp. 479–495.

[206] D. Goodin. "Don't count on STARTTLS to automatically encrypt your sensitive e-mails". In: *Ars Technica* (Oct. 30, 2015). URL: https://arstechnica.com/information-technology/2015/10/dont-count-on-starttls-to-automatically-encrypt-your-sensitive-e-mails/.

[207] D. Goodin. "HTTPS-crippling attack threatens tens of thousands of Web and mail servers". In: *Ars Technica* (May 20, 2015). URL: https://arstechnica.com/information-technology/2015/05/https-crippling-attack-threatens-tens-of-thousands-of-web-and-mail-servers/.

[208] D. Goodin. *More than 11 million HTTPS websites imperiled by new decryption attack*. Mar. 1, 2016. URL: https://arstechnica.com/information-technology/2016/03/more-than-13-million-https-websites-imperiled-by-new-decryption-attack/.

[209] T. Groß. "Statistical Reliability of 10 Years of Cyber Security User Studies". In: *10th International Workshop on Socio-Technical Aspects in Security and Trust*. STAST '20. 2020, pp. 171–190. DOI: 10.1007/978-3-030-79318-0_10.

[210] G. Gu. *Computer Security Conference Ranking and Statistic*. 2022. URL: https://people.engr.tamu.edu/guofei/sec_conf_stat.htm.

[211] M. Gundersen. *Dette er de norske nettsidene som sporer deg mest*. Norwegian. NRKbeta. May 28, 2019. URL: https://nrkbeta.no/2019/05/28/dette-er-de-norske-nettsidene-som-sporer-deg-mest/.

[212] J. Guynn. "Facebook takes down ads mentioning African-Americans and Hispanics, calling them political". In: *USA TODAY* (Oct. 17, 2018). URL: https://www.usatoday.com/story/news/2018/10/17/facebook-labels-african-american-hispanic-mexican-ads-political/1608841002/.

[213] M. Guzdial. *Why I Don't Recommend CSRankings.org: Know the Values You are Ranking On*. BLOG@CACM. Oct. 18, 2020. URL: https://cacm.acm.org/blogs/blog-cacm/248078-why-i-dont-recommend-csrankingsorg-know-the-values-you-are-ranking-on/fulltext.

[214] T. S. Guzella and W. M. Caminhas. "A review of machine learning approaches to Spam filtering". In: *Expert Systems with Applications* 36.7 (2009), pp. 10206–10222. DOI: 10.1016/j.eswa.2009.02.037.

[215] Z. Gyöngyi, P. Berkhin, H. Garcia-Molina, and J. Pedersen. "Link Spam Detection Based on Mass Estimation". In: *32nd International Conference on Very Large Data Bases*. VLDB '06. 2006, pp. 439–450.

[216] S. Hajian, F. Bonchi, and C. Castillo. "Algorithmic Bias: From Discrimination Discovery to Fairness-Aware Data Mining". In: *22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. KDD '16. 2016, pp. 2125–2126. DOI: 10.1145/2939672.2945386.

[217] A. Halavais and D. Lackaff. "An Analysis of Topical Coverage of Wikipedia". In: *Journal of Computer-Mediated Communication* 13.2 (2008), pp. 429–440. DOI: 10.1111/j.1083-6101.2008.00403.x.

[218] J. van der Ham and R. van Rijswijk-Deij. "Ethics and Internet Measurements". In: *Journal of Cyber Security and Mobility* 5.4 (2017), pp. 287–308. DOI: 10.13052/jcsm2245-1439.543.

[219] P. Hamm, D. Harborth, and S. Pape. "A Systematic Analysis of User Evaluations in Security Research". In: *14th International Conference on Availability, Reliability and Security*. ARES '19. 2019. DOI: 10.1145/3339252.3340339.

[220] F. Hantke, S. Calzavara, M. Wilhelm, A. Rabitti, and B. Stock. "You Call This Archaeology? Evaluating Web Archives for Reproducible Web Security Measurements". In: *2023 ACM Conference on Computer and Communications Security*. CCS '23. 2023.

[221] S. Hao, A. Kantchelian, B. Miller, V. Paxson, and N. Feamster. "PREDATOR: Proactive Recognition and Elimination of Domain Abuse at Time-Of-Registration". In: *2016 ACM SIGSAC Conference on Computer and Communications Security*. CCS '16. 2016, pp. 1568–1579. DOI: 10.1145/2976749.2978317.

[222] S. Hao, M. Thomas, V. Paxson, N. Feamster, C. Kreibich, C. Grier, and S. Hollenbeck. "Understanding the Domain Registration Behavior of Spammers". In: *13th Internet Measurement Conference*. IMC '13. 2013, pp. 63–76. DOI: 10.1145/2504730.2504753.

[223] M. Haroon, A. Chhabra, X. Liu, P. Mohapatra, Z. Shafiq, and M. Wojcieszak. *YouTube, The Great Radicalizer? Auditing and Mitigating Ideological Biases in YouTube Recommendations*. 2022. DOI: 10.48550/arxiv.2203.10666. arXiv: 2203.10666.

[224] M. Heinemeyer. *How malware abused Sixt.com and Breitling.com for covert Command & Control communication*. Darktrace. Mar. 7, 2018. URL: https://www.darktrace.com/en/blog/how-malware-abused-sixt-com-and-breitling-com-for-covert-command-control-communication/.

[225] S. Helme. *Top 1 Million Analysis - September 2019*. Oct. 3, 2019. URL: https://scotthelme.co.uk/top-1-million-analysis-september-2019/.

[226] E. Hjelmvik. *Domain Whitelist Benchmark: Alexa vs Umbrella*. NETRESEC. Apr. 3, 2017. URL: https://www.netresec.com/?page=Blog&month=2017-04&post=Domain-Whitelist-Benchmark%3A-Alexa-vs-Umbrella.

[227] J. van Hoboken, N. Appelman, R. Ó Fathaigh, P. Leerssen, T. McGonagle, N. van Eijk, and N. Helberger. *The legal framework on the dissemination of disinformation through Internet services and the regulation of political advertising*. Institute for Information Law, University of Amsterdam, Dec. 2019. URL: https://www.ivir.nl/publicaties/download/Report_Disinformation_Dec2019-1.pdf.

[228] P. Hoffman. *Collecting "Typical" Domain Names for Web Servers*. OCTO-023. ICANN Office of the Chief Technology Officer, Feb. 24, 2021. URL: https://www.icann.org/en/system/files/files/octo-023-24feb21-en.pdf.

[229] P. Hoffman. *DNSSEC Algorithm Use in 2022*. OCTO-033. ICANN Office of the Chief Technology Officer, Apr. 4, 2022. URL: https://www.icann.org/en/system/files/files/octo-033-04apr22-en.pdf.

[230] T. Holz, C. Gorecki, K. Rieck, and F. C. Freiling. "Measuring and Detecting Fast-Flux Service Networks". In: *15th Annual Network and Distributed System Security Symposium*. NDSS '08. 2008.

[231] J. Horwitz. "Facebook Seeks Shutdown of NYU Research Project Into Political Ad Targeting". In: *The Wall Street Journal* (Oct. 23, 2020). URL: https://www.wsj.com/articles/facebook-seeks-shutdown-of-nyu-research-project-into-political-ad-targeting-11603488533.

[232] *How Ads About Social Issues, Elections or Politics Are Reviewed (With Examples)*. Facebook Business Help Center. Mar. 3, 2021. URL: https://www.facebook.com/business/help/313752069181919.

[233] *How Disclaimers Work for Ads About Social Issues, Elections or Politics*. Facebook Business Help Center. Aug. 19, 2020. URL: https://www.facebook.com/business/help/198009284345835.

[234] *How do I choose to see fewer ads about social issues, elections or politics on Facebook?* Facebook Help Center. 2021. URL: https://www.facebook.com/help/595432167810439.

[235] P. N. Howard. "A Way to Detect the Next Russian Misinformation Campaign". In: *The New York Times* (Mar. 27, 2019). URL: https://www.nytimes.com/2019/03/27/opinion/russia-elections-facebook.html.

[236] D. Hubbard. *Cisco Umbrella 1 Million*. Dec. 14, 2016. URL: https://umbrella.cisco.com/blog/2016/12/14/cisco-umbrella-1-million/.

[237] T. Hunt. *Why No HTTPS?* Aug. 12, 2021. URL: https://whynohttps.com/.

[238] E. Hussein, P. Juneja, and T. Mitra. "Measuring Misinformation in Video Search Platforms: An Audit Study on YouTube". In: *Proceedings of the ACM on Human-Computer Interaction* 4.CSCW1 (May 2020). DOI: 10.1145/3392854.

[239] A. Hutchings, R. Clayton, and R. Anderson. "Taking down websites to prevent crime". In: *2016 APWG Symposium on Electronic Crime Research*. eCrime '16. 2016. DOI: 10.1109/ECRIME.2016.7487947.

[240] IBM Security. *IBM X-Force Exchange. Frequently Asked Questions*. URL: https://exchange.xforce.ibmcloud.com/faq.

[241] B. Imana, A. Korolova, and J. Heidemann. "Auditing for Discrimination in Algorithms Delivering Job Ads". In: *The Web Conference 2021*. WWW '21. 2021, pp. 3767–3778. DOI: 10.1145/3442381.3450077.

[242] B. Imana, A. Korolova, and J. Heidemann. *Having your Privacy Cake and Eating it Too: Platform-supported Auditing of Social Media Algorithms for Public Interest*. July 2022. DOI: 10.48550/arXiv.2207.08773. arXiv: 2207.08773 [cs.CY].

[243] *Inquiry into and report on all aspects of the conduct of the 2019 Federal Election and matters related thereto*. Submission to the Joint Standing Committee on Electoral Matters 140.1. Facebook, Sept. 16, 2020. URL: https://www.aph.gov.au/DocumentStore.ashx?id=acd7119e-7cc7-474f-ac60-6af6c5c3d933&subId=671217.

[244] Internet Corporation for Assigned Names and Numbers. *How long does a registration last? Can it be renewed?* Feb. 25, 2012. URL: https://www.icann.org/resources/pages/faqs-84-2012-02-25-en%5C#7.

[245] Internet Corporation for Assigned Names and Numbers. *Temporary Specification for gTLD Registration Data*. Internet Corporation for Assigned Names and Numbers. May 2018. URL: https://www.icann.org/resources/pages/gtld-registration-data-specs-en.

[246] *Internet Society Pulse – Partners*. Internet Society. 2021. URL: https://pulse.internetsociety.org/partners.

[247] L. Invernizzi, K. Thomas, A. Kapravelos, O. Comanescu, J. Picod, and E. Bursztein. "Cloak of Visibility: Detecting When Machines Browse a Different Web". In: *2016 IEEE Symposium on Security and Privacy*. SP '16. 2016, pp. 743–758. DOI: 10.1109/SP.2016.50.

[248] J. P. A. Ioannidis. "Why Most Published Research Findings Are False". In: *PLoS Medicine* 2.8 (Aug. 2005), e124. DOI: 10.1371/journal.pmed.0020124.

[249] J. P. A. Ioannidis, D. Fanelli, D. D. Dunne, and S. N. Goodman. "Meta-research: Evaluation and Improvement of Research Methods and Practices". In: *PLOS Biology* 13.10 (Oct. 2015), e1002264. DOI: 10.1371/journal.pbio.1002264.

[250] K. Iwańska, K. Szymielewicz, D. Batorski, M. Baranowski, J. Krawiec, K. Izdebski, and D. Macyszyn. *Who (really) targets you? Facebook in Polish election campaigns*. Panoptykon Foundation, ePaństwo Foundation, SmartNet Research & Solutions, Mar. 17, 2020. URL: https://panoptykon.org/political-ads-report.

[251] A. M. Jamison, D. A. Broniatowski, M. Dredze, Z. Wood-Doughty, D. Khan, and S. C. Quinn. "Vaccine-related advertising in the Facebook Ad Archive". In: *Vaccine* 38.3 (2020), pp. 512–520. DOI: 10.1016/j.vaccine.2019.10.066.

[252] J. Jaursch. *Defining Online Political Advertising. How Difficulties in Delineating Paid Political Communication Can Be Addressed*. Stiftung Neue Verantwortung, Nov. 2020. URL: https://www.stiftung-nv.de/sites/default/files/snv_definingpoliticalads.pdf.

[253] S. Jeffers. *How to take a "gold standard" approach to political advertising transparency and policy*. Who Targets Me. Oct. 28, 2020. URL: https://whotargets.me/how-to-take-a-gold-standard-approach-to-political-advertising-transparency-and-policy/.

[254] D. Jones. *Majestic Million CSV now free for all, daily.* Oct. 1, 2012. URL: https://blog.majestic.com/development/majestic-million-csv-daily/.

[255] B. Juba, C. Musco, F. Long, S. Sidiroglou-Douskos, and M. Rinard. "Principled Sampling for Anomaly Detection". In: *22nd Annual Network and Distributed System Security Symposium*. NDSS '15. 2015. DOI: 10.14722/ndss.2015.23268.

[256] J. Jueckstock, S. Sarker, P. Snyder, A. Beggs, P. Papadopoulos, M. Varvello, B. Livshits, and A. Kapravelos. "Towards Realistic and Reproducible Web Crawl Measurements". In: *The Web Conference 2021*. WWW '21. 2021, pp. 80–91. DOI: 10.1145/3442381.3450050.

[257] M. Karami, Y. Park, and D. McCoy. "Stress Testing the Booters: Understanding and Undermining the Business of DDoS Services". In: *25th International Conference on World Wide Web*. WWW '16. 2016, pp. 1033–1043. DOI: 10.1145/2872427.2883004.

[258] M. Kaur, M. van Eeten, M. Janssen, K. Borgolte, and T. Fiebig. *Human Factors in Security Research: Lessons Learned from 2008-2018*. 2021. DOI: 10.48550/arxiv.2103.13287. arXiv: 2103.13287.

[259] S. Kemp. *Digital 2021: Global Overview Report*. Hootsuite & We Are Social, Jan. 27, 2021. URL: https://datareportal.com/reports/digital-2021-global-overview-report.

[260] L. Keselman. "Venue Analytics: A Simple Alternative to Citation-Based Metrics". In: *2019 ACM/IEEE Joint Conference on Digital Libraries*. JCDL '19. 2019, pp. 315–324. DOI: 10.1109/JCDL.2019.00052.

[261] S. Keshav. "Editor's Message". In: *ACM SIGCOMM Computer Communication Review* 41.3 (July 2011). DOI: 10.1145/2002250.

[262] N. Kheir, F. Tran, P. Caron, and N. Deschamps. "Mentor: Positive DNS Reputation to Skim-Off Benign Domains in Botnet C&C Blacklists". In: *29th IFIP International Information Security and Privacy Conference*. SEC '14. 2014, pp. 1–14. DOI: 10.1007/978-3-642-55415-5_1.

[263] E. Kidmose, E. Lansing, S. Brandbyge, and J. M. Pedersen. "Detection of Malicious and Abusive Domain Names". In: *1st International Conference on Data Intelligence and Security*. ICDIS '18. 2018, pp. 49–56. DOI: 10.1109/ICDIS.2018.00015.

[264] J. Killock. *Facebook Don't Want You To Know How Their Algorithm Works*. Open Rights Group. Mar. 29, 2018. URL: https://www.openrightsgroup.org/blog/facebook-dont-want-you-to-know-how-their-algorithm-works-2/.

[265] P. T. Kim. "Auditing Algorithms for Discrimination". In: *University of Pennsylvania Law Review Online* 166 (2017), pp. 189–204. URL: https://scholarship.law.upenn.edu/penn_law_review_online/vol166/iss1/10/.

[266] J. King. *Bringing More Transparency to Social Issue, Electoral and Political Ads*. Meta. May 22, 2022. URL: https://www.facebook.com/business/news/transparency-social-issue-electoral-political-ads.

[267] K. Kopel. "Operation Seizing Our Sites: How the Federal Government is Taking Domain Names Without Prior Notice". In: *Berkeley Technology Law Journal* 28.4 (2013): *Annual Review 2013*, pp. 859–900. DOI: 10.15779/Z384Q3M.

[268] M. Korczyński, S. Tajalizadehkhoob, A. Noroozian, M. Wullink, C. Hesselman, and M. van Eeten. "Reputation Metrics Design to Improve Intermediary Incentives for Security of TLDs". In: *2017 IEEE European Symposium on Security and Privacy*. EuroS&P '17. 2017, pp. 579–594. DOI: 10.1109/EuroSP.2017.15.

[269] M. Korczyński, M. Wullink, S. Tajalizadehkhoob, G. C. M. Moura, A. Noroozian, D. Bagley, and C. Hesselman. "Cybercrime After the Sunrise: A Statistical Analysis of DNS Abuse in New gTLDs". In: *2018 ACM Asia Conference on Computer and Communications Security*. ASIACCS '18. 2018, pp. 609–623. DOI: 10.1145/3196494.3196548.

[270] A. Kountouras, P. Kintis, C. Lever, Y. Chen, Y. Nadji, D. Dagon, M. Antonakakis, and R. Joffe. "Enabling Network Security Through Active DNS Datasets". In: *Research in Attacks, Intrusions, and Defenses*. RAID '16. 2016, pp. 188–208. DOI: 10.1007/978-3-319-45719-2_9.

[271] E. van der Kouwe, G. Heiser, D. Andriesse, H. Bos, and C. Giuffrida. "SoK: Benchmarking Flaws in Systems Security". In: *2019 IEEE European Symposium on Security and Privacy*. EuroS&P '19. 2019, pp. 310–325. DOI: 10.1109/EuroSP.2019.00031.

[272] E. Kovacs. "Amazon's Shuttering of Alexa Ranking Service Hits Cybersecurity Industry". In: *SecurityWeek* (May 6, 2022). URL: https://www.securityweek.com/impact-alexa-ranking-service-shutdown-cybersecurity-industry.

[273] T. Krause, R. Ernst, B. Klaer, I. Hacker, and M. Henze. "Cybersecurity in Power Grids: Challenges and Opportunities". In: *Sensors* 21.18, 6225 (2021). DOI: 10.3390/s21186225.

[274] D. Kreiss and B. Barrett. "Democratic Tradeoffs: Platforms and Political Advertising". In: *Ohio State Technology Law Journal* 16.2 (2020), pp. 493–519.

[275] K. Krippendorff. *Content analysis: an introduction to its methodology*. 4th ed. Sage, 2018. ISBN: 9781506395661.

[276] S. Krishnan, T. Taylor, F. Monrose, and J. McHugh. "Crossing the threshold: Detecting network malfeasance via sequential hypothesis testing". In: *43rd Annual IEEE/IFIP International Conference on Dependable Systems and Networks*. DSN '13. 2013. DOI: 10.1109/DSN.2013.6575364.

[277] K. Krol, J. M. Spring, S. Parkin, and M. A. Sasse. "Towards Robust Experimental Design for User Studies in Security and Privacy". In: *2016 Learning from Authoritative Security Experiment Results Workshop*. LASER '16. 2016, pp. 21–31. URL: https://www.usenix.org/conference/laser2016/program/presentation/krol.

[278] J. A. Kroll, J. Huey, S. Barocas, E. W. Felten, J. R. Reidenberg, D. G. Robinson, and H. Yu. "Accountable Algorithms". In: *University of Pennsylvania Law Review* 165.3 (2017), pp. 633–705. URL: https://scholarship.law.upenn.edu/penn_law_review/vol165/iss3/3.

[279] B. Krumnow, H. Jonker, and S. Karsch. "How Gullible Are Web Measurement Tools? A Case Study Analysing and Strengthening OpenWPM's Reliability". In: *18th International Conference on Emerging Networking EXperiments and Technologies*. CoNEXT '22. 2022, pp. 171–186. DOI: 10.1145/3555050.3569131.

[280] M. Kührer, C. Rossow, and T. Holz. "Paint It Black: Evaluating the Effectiveness of Malware Blacklists". In: *17th International Symposium on Research in Attacks, Intrusions and Defenses*. RAID '14. 2014, pp. 1–21. DOI: 10.1007/978-3-319-11379-1_1.

[281] D. Kumar, Z. Ma, Z. Durumeric, A. Mirian, J. Mason, J. A. Halderman, and M. Bailey. "Security Challenges in an Increasingly Tangled Web". In: *26th International Conference on World Wide Web*. WWW '17. 2017, pp. 677–684. DOI: 10.1145/3038912.3052686.

[282] P. Küngas, S. Karus, S. Vakulenko, M. Dumas, C. Parra, and F. Casati. "Reverse-engineering conference rankings: what does it take to make a reputable conference?" In: *Scientometrics* 96.2 (Jan. 2013), pp. 651–665. DOI: 10.1007/s11192-012-0938-8.

[283] A. Lambrecht and C. Tucker. "Algorithmic Bias? An Empirical Study of Apparent Gender-Based Discrimination in the Display of STEM Career Ads". In: *Management Science* 65.7 (July 2019), pp. 2966–2981. DOI: 10.1287/mnsc.2018.3093.

[284] I. Lapowsky. "How right-wing websites are getting around Facebook's political ad ban". In: *Protocol* (Dec. 11, 2020). URL: https://www.protocol.com/facebook-political-ad-ban-news.

[285] I. Lapowsky. "The FTC hits back at Facebook after it shut down NYU research". In: *Protocol* (Aug. 5, 2021). URL: https://www.protocol.com/ftc-zuckerberg-nyu-letter.

[286] B. Laurie, A. Langley, and E. Kasper. *Certificate Transparency*. RFC 6962. RFC Editor, June 2013.

[287] V. Le Pochat, L. Edelson, T. Van Goethem, W. Joosen, D. McCoy, and T. Lauinger. "An Audit of Facebook's Political Ad Policy Enforcement". In: *31st USENIX Security Symposium*. USENIX Security '22. 2022, pp. 607–624.

[288] V. Le Pochat, T. Van Goethem, and W. Joosen. "A Smörgåsbord of Typos: Exploring International Keyboard Layout Typosquatting". In: *2019 IEEE Security and Privacy Workshops*. SPW '19. 2019, pp. 187–192. DOI: 10.1109/SPW.2019.00043.

[289] V. Le Pochat, T. Van Goethem, and W. Joosen. "Evaluating the Long-term Effects of Parameters on the Characteristics of the Tranco Top Sites Ranking". In: *12th USENIX Workshop on Cyber Security Experimentation and Test*. CSET '19. 2019. URL: https://www.usenix.org/conference/cset19/presentation/lepochat.

[290] V. Le Pochat, T. Van Goethem, and W. Joosen. "Funny Accents: Exploring Genuine Interest in Internationalized Domain Names". In: *20th Passive and Active Measurement Conference*. PAM '19. 2019, pp. 178–194. DOI: 10.1007/978-3-030-15986-3_12.

[291] V. Le Pochat, T. Van Goethem, S. Tajalizadehkhoob, M. Korczyński, and W. Joosen. "Tranco: A Research-Oriented Top Sites Ranking Hardened Against Manipulation". In: *26th Annual Network and Distributed System Security Symposium*. NDSS '19. 2019. DOI: 10.14722/ndss.2019.23386.

[292] V. Le Pochat, T. Van hamme, S. Maroofi, T. Van Goethem, D. Preuveneers, A. Duda, W. Joosen, and M. Korczyński. "A Practical Approach for Taking Down Avalanche Botnets Under Real-World Constraints". In: *27th Annual Network and Distributed System Security Symposium*. NDSS '20. 2020. DOI: 10.14722/ndss.2020.24161.

[293] H. Ledford. "Researchers scramble as Twitter plans to end free data access". In: *Nature* 614.7949 (Feb. 2023), pp. 602–603. DOI: 10.1038/d41586-023-00460-z.

[294] E. Lee. *The Toxic Culture of Rejection in Computer Science*. ACM SIGBED. Aug. 22, 2022. URL: https://sigbed.org/2022/08/22/the-toxic-culture-of-rejection-in-computer-science/.

[295] P. Leerssen, J. Ausloos, B. Zarouali, N. Helberger, and C. H. de Vreese. "Platform ad archives: promises and pitfalls". In: *Internet Policy Review* 8.4 (Oct. 2019). DOI: 10.14763/2019.4.1421.

[296] J. Lehmann, R. Isele, M. Jakob, A. Jentzsch, D. Kontokostas, P. N. Mendes, S. Hellmann, M. Morsey, P. van Kleef, S. Auer, and C. Bizer. "DBpedia – A large-scale, multilingual knowledge base extracted from Wikipedia". In: *Semantic Web* 6.2 (2015), pp. 167–195. DOI: 10.3233/SW-140134.

[297] R. Lerman and C. Zakrzewski. "Facebook ban on new political ads starts off with major hiccups". In: *The Washington Post* (Oct. 27, 2020). URL: https://www.washingtonpost.com/technology/2020/10/27/facebook-ban-new-political-ads/.

[298] C. Lever, R. Walls, Y. Nadji, D. Dagon, P. McDaniel, and M. Antonakakis. "Domain-Z: 28 Registrations Later. Measuring the Exploitation of Residual Trust in Domains". In: *2016 IEEE Symposium on Security and Privacy*. SP '16. 2016, pp. 691–706. DOI: 10.1109/SP.2016.47.

[299] T. Libert. "An Automated Approach to Auditing Disclosure of Third-Party Data Collection in Website Privacy Policies". In: *2018 World Wide Web Conference*. WWW '18. 2018, pp. 207–216. DOI: 10.1145/3178876.3186087.

[300] T. Libert. "Exposing the Hidden Web: An Analysis of Third-Party HTTP Requests on 1 Million Websites". In: *International Journal of Communication* 9 (Oct. 2015), pp. 3544–3561.

[301] lift_ticket83. *Reddit Data API Update: Changes to Pushshift Access*. Reddit. May 2, 2023. URL: https://www.reddit.com/r/modnews/comments/134tjpe/reddit_data_api_update_changes_to_pushshift_access/.

[302] P. Lison and V. Mavroeidis. "Neural reputation models learned from passive DNS data". In: *2017 IEEE International Conference on Big Data*. Big Data '17. 2017, pp. 3662–3671. DOI: 10.1109/BigData.2017.8258361.

[303] D. Liu and R. Duan. *Dangling Domains: Security Threats, Detection and Prevalence*. Palo Alto Networks – Unit 42. Sept. 16, 2021. URL: https://unit42.paloaltonetworks.com/dangling-domains/.

[304] L. Liu, G. Engelen, T. Lynar, D. Essam, and W. Joosen. "Error Prevalence in NIDS datasets: A Case Study on CIC-IDS-2017 and CSE-CIC-IDS-2018". In: *2022 IEEE Conference on Communications and Network Security*. CNS '22. 2022.

[305] S. Liu, I. Foster, S. Savage, G. M. Voelker, and L. K. Saul. "Who is .com?: Learning to Parse WHOIS Records". In: *15th Internet Measurement Conference*. 2015, pp. 369–380. DOI: 10.1145/2815675.2815693.

[306] E. Llansó, J. van Hoboken, P. Leerssen, and J. Harambam. *Artificial Intelligence, Content Moderation, and Freedom of Expression*. Transatlantic Working Group, 2020. URL: https://www.ivir.nl/publicaties/download/AI-Llanso-Van-Hoboken-Feb-2020.pdf.

[307] B. W. N. Lo and R. S. Sedhain. "How reliable are website rankings? Implications for e-business advertising and Internet search". In: *Issues in Information Systems* VII.2 (2006), pp. 233–238.

[308] T. Longstaff, D. Balenson, and M. Matties. "Barriers to Science in Security". In: *26th Annual Computer Security Applications Conference*. ACSAC '10. 2010, pp. 127–129. DOI: 10.1145/1920261.1920281.

[309] P. Lowe. *DNSFilter Top Domains*. DNSFilter, 2023. URL: https://github.com/DNSFilter/topdomains.

[310] O. Lystrup. *Cisco Umbrella Releases Free Top 1 Million Sites List*. Dec. 20, 2016. URL: https://medium.com/cisco-shifted/cisco-umbrella-releases-free-top-1-million-sites-list-8497fba58efe.

[311] J. Ma, L. K. Saul, S. Savage, and G. M. Voelker. "Beyond Blacklists: Learning to Detect Malicious Web Sites from Suspicious URLs". In: *15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. KDD '09. 2009, pp. 1245–1254. DOI: 10.1145/1557019.1557153.

[312] L. Machlica, K. Bartos, and M. Sofka. *Learning detectors of malicious web requests for intrusion detection in network traffic*. Feb. 2017. arXiv: 1702.02530 [stat.ML].

[313] K. Macnish and J. van der Ham. "Ethics in cybersecurity research and practice". In: *Technology in Society* 63, 101382 (Nov. 2020). DOI: 10.1016/j.techsoc.2020.101382.

[314] S. Maher. *There's a Fix to Disinformation: Make Social Media Algorithms Transparent*. Centre for International Governance Innovation. Mar. 15, 2022. URL: https://www.cigionline.org/articles/algorithmic-transparency-should-be-considered-part-of-national-security/.

[315] Majestic-12 Ltd. *Frequently Asked Questions*. URL: https://majestic.com/support/faq.

[316] Majestic-12 Ltd. *Majestic launch a Bigger Fresh Index*. Apr. 12, 2018. URL: https://blog.majestic.com/company/majestic-launch-a-bigger-fresh-index/.

[317] A. Mak. "Facebook Thought an Ad From Bush's Baked Beans Was "Political" and Removed It". In: *Slate* (May 30, 2018). URL: https://slate.com/technology/2018/05/bushs-baked-beans-fell-victim-to-facebooks-political-ads-system.html.

[318] S. Maroofi, M. Korczyński, C. Hesselman, B. Ampeau, and A. Duda. "COMAR: Classification of Compromised versus Maliciously Registered Domains". In: *2020 IEEE European Symposium on Security and Privacy*. EuroS&P '20. 2020, pp. 607–623. DOI: 10.1109/EuroSP48549.2020.00045.

[319] F. Marquardt and C. Schmidt. "Don't Stop at the Top: Using Certificate Transparency Logs to Extend Domain Lists for Web Security Studies". In: *45th IEEE Conference on Local Computer Networks*. LCN '20. 2020, pp. 409–412. DOI: 10.1109/LCN48667.2020.9314793.

[320] C. Martinho and S. Zejnilovic. *Goodbye, Alexa. Hello, Cloudflare Radar Domain Rankings*. The Cloudflare Blog. Sept. 30, 2022. URL: https://blog.cloudflare.com/radar-domain-rankings/.

[321] J. N. Matias, A. Hounsel, and N. Feamster. "Software-Supported Audits of Decision-Making Systems: Testing Google and Facebook's Political Advertising Policies". In: *Proceedings of the ACM on Human-Computer Interaction* 6.CSCW1, 118 (Apr. 2022). DOI: 10.1145/3512965.

[322] J. N. Matias, A. Hounsel, and M. Hopkins. "We Tested Facebook's Ad Screeners and Some Were Too Strict". In: *The Atlantic* (Nov. 2, 2018). URL: https://www.theatlantic.com/technology/archive/2018/11/do-big-social-media-platforms-have-effective-ad-policies/574609/.

[323] S. Mattu and A. Sankin. "How We Built a Real-time Privacy Inspector". In: *The Markup* (Sept. 22, 2020). URL: https://themarkup.org/blacklight/2020/09/22/how-we-built-a-real-time-privacy-inspector.

[324] *Media Bias/Fact Check*. Media Bias Fact Check, LLC. 2021. URL: https://mediabiasfactcheck.com/.

[325] J. B. Merrill and A. Tobin. "Facebook Moves to Block Ad Transparency Tools — Including Ours". In: *ProPublica* (Jan. 28, 2019). URL: https://www.propublica.org/article/facebook-blocks-ad-transparency-tools.

[326] X. Mertens. *Whitelists: The Holy Grail of Attackers*. SANS Internet Storm Center. Apr. 5, 2017. URL: https://isc.sans.edu/forums/diary/Whitelists+The+Holy+Grail+of+Attackers/22262/.

[327] L. B. Metcalf, D. Ruef, and J. M. Spring. "Open-source Measurement of Fast-flux Networks While Considering Domain-name Parking". In: *2017 Learning from Authoritative Security Experiment Results Workshop*. LASER '17. 2017, pp. 13–24.

[328] A. Metwally, D. Agrawal, A. E. Abbad, and Q. Zheng. "On Hit Inflation Techniques and Detection in Streams of Web Advertising Networks". In: *27th International Conference on Distributed Computing Systems*. ICDCS '07. 2007, 52. DOI: 10.1109/ICDCS.2007.124.

[329] R. Meusel, S. Vigna, O. Lehmberg, and C. Bizer. "The Graph Structure in the Web – Analyzed on Different Aggregation Levels". In: *The Journal of Web Science* 1.1 (2015), pp. 33–47. DOI: 10.1561/106.00000003.

[330] Z. Meyers. *Will the Digital Services Act save Europe from disinformation?* Centre for European Reform. Apr. 21, 2022. URL: https://www.cer.eu/insights/will-digital-services-act-save-europe-disinformation.

[331] A. Miagkov, A. Arrieta, and B. Cyphers. *Giving Privacy Badger a Jump Start*. Electronic Frontier Foundation. Aug. 22, 2018. URL: https://www.eff.org/deeplinks/2018/08/giving-privacy-badger-jump-start.

[332] B. Miller et al. "Reviewer Integration and Performance Measurement for Malware Detection". In: *13th International Conference on Detection of Intrusions and Malware, and Vulnerability Assessment*. DIMVA '16. 2016, pp. 122–141. DOI: 10.1007/978-3-319-40667-1_7.

[333] C. Miller. "Facebook, It's Time to Put the Rules in One Place". In: *Lawfare* (Mar. 5, 2021). URL: https://www.lawfareblog.com/facebook-its-time-put-rules-one-place.

[334] J. Mink, H. Benkraouda, L. Yang, A. Ciptadi, A. Ahmadzadeh, D. Votipka, and G. Wang. "Everybody's Got ML, Tell Me What Else You Have: Practitioners' Perception of ML-Based Security Tools and Explanations". In: *44th IEEE Symposium on Security and Privacy*. SP '23. 2023.

[335] R. Mir and C. Doctorow. *Facebook's Attack on Research is Everyone's Problem*. Electronic Frontier Foundation. Aug. 12, 2021. URL: https://www.eff.org/deeplinks/2021/08/facebooks-attack-research-everyones-problem.

[336] Y. Mirsky, T. Doitshman, Y. Elovici, and A. Shabtai. "Kitsune: An Ensemble of Autoencoders for Online Network Intrusion Detection". In: *2018 Network and Distributed System Security Symposium*. 2018. DOI: 10.14722/ndss.2018.23204.

[337] *misp-warninglists*. MISP, 2022. URL: https://github.com/MISP/misp-warninglists.

[338] B. Mittelstadt, C. Russell, and S. Wachter. "Explaining Explanations in AI". In: *2nd Conference on Fairness, Accountability, and Transparency*. FAT* '19. 2019, pp. 279–288. DOI: 10.1145/3287560.3287574.

[339] J. C. Mogul. "Towards More Constructive Reviewing of SIGCOMM Papers". In: *ACM SIGCOMM Computer Communication Review* 43.3 (July 2013), pp. 90–94. DOI: 10.1145/2500098.2500112.

[340] J. C. Mogul and T. Anderson. "Open issues in organizing computer systems conferences". In: *ACM SIGCOMM Computer Communication Review* 38.3 (July 2008), pp. 93–102. DOI: 10.1145/1384609.1384623.

[341] J. Mökander, J. Morley, M. Taddeo, and L. Floridi. "Ethics-Based Auditing of Automated Decision-Making Systems: Nature, Scope, and Limitations". In: *Science and Engineering Ethics* 27.4, 44 (July 2021). DOI: 10.1007/s11948-021-00319-4.

[342] B. Montgomery. "Facebook Axed Pro-Vaccine Ads, Let Anti-Vaxxer Conspiracies Slip Through". In: *The Daily Beast* (Oct. 25, 2019). URL: https://www.thedailybeast.com/facebook-axed-pro-vaccine-ads-from-hospitals-and-health-orgs-let-anti-vaxxer-ads-slip-through.

[343] B. Montgomery. "Facebook's Moderators Took Down the Tech Giant's Own Pro-Equality Ads". In: *The Daily Beast* (Mar. 10, 2021). URL: https://www.thedailybeast.com/facebooks-moderation-system-took-down-the-tech-giants-own-black-history-month-ads.

[344] T. Moore, E. Kenneally, M. Collett, and P. Thapa. "Valuing cybersecurity research datasets". In: *18th Workshop on the Economics of Information Security*. WEIS '19. 2019. URL: https://tylermoore.utulsa.edu/weis19data.pdf.

[345] B. Morton. *Protect Your Domain with CT Search*. Oct. 2016. URL: https://www.entrustdatacard.com/blog/2016/october/protect-your-domain-with-ct-search.

[346] R. Moss and E. Krueger. *We Bid Goodbye to Alexa Rankings, and Measure Its Contribution to the Tranco List (Pre-May)*. DeepSee. May 25, 2022. URL: https://deepsee.io/blog/we-bid-goodbye-to-alexa-rankings-and-measure-its-contribution-to-the-tranco-list-pre-may.

[347] G. C. M. Moura, M. Müller, M. Davids, M. Wullink, and C. Hesselman. "Domain names abuse and TLDs: From monetization towards mitigation". In: *2017 IFIP/IEEE Symposium on Integrated Network and Service Management*. IM '17. 2017, pp. 1077–1082. DOI: 10.23919/INM.2017.7987441.

[348] M. Mowbray and J. Hagen. "Finding Domain-Generation Algorithms by Looking at Length Distribution". In: *2014 IEEE International Symposium on Software Reliability Engineering Workshops*. ISSREW' 14. 2014, pp. 395–400. DOI: 10.1109/ISSREW.2014.20.

[349] Mozilla Foundation. *Public Suffix List*. 2019. URL: https://publicsuffix.org/.

[350] M. R. Munafò, B. A. Nosek, D. V. M. Bishop, K. S. Button, C. D. Chambers, N. P. du Sert, U. Simonsohn, E.-J. Wagenmakers, J. J. Ware, and J. P. A. Ioannidis. "A manifesto for reproducible science". In: *Nature Human Behaviour* 1.1 (Jan. 2017). DOI: 10.1038/s41562-016-0021.

[351] M. Musch, R. Kirchner, M. Boll, and M. Johns. "Server-Side Browsers: Exploring the Web's Hidden Attack Surface". In: *2022 ACM Asia Conference on Computer and Communications Security*. ASIA CCS '22. 2022, pp. 1168–1181. DOI: 10.1145/3488932.3517414.

[352] J. Naab, P. Sattler, J. Jelten, O. Gasser, and G. Carle. "Prefix Top Lists: Gaining Insights with Prefixes from Domain-based Top Lists on DNS Deployment". In: *19th Internet Measurement Conference*. IMC '19. 2019, pp. 351–357. DOI: 10.1145/3355369.3355598.

[353] Y. Nadji, M. Antonakakis, R. Perdisci, D. Dagon, and W. Lee. "Beheading Hydras: Performing Effective Botnet Takedowns". In: *2013 ACM SIGSAC Conference on Computer and Communications Security*. CCS '13. 2013, pp. 121–132. DOI: 10.1145/2508859.2516749.

[354] P. M. Napoli, P. J. Lavrakas, and M. Callegaro. "Internet and mobile ratings panels". In: *Online Panel Research: A Data Quality Perspective*. Wiley-Blackwell, 2014. Chap. 17, pp. 387–407. ISBN: 9781118763520. DOI: 10.1002/9781118763520.ch17.

[355] A. Narayanan and K. Lee. "Security Policy Audits: Why and How". In: *IEEE Security & Privacy* 21.2 (2023), pp. 77–81. DOI: 10.1109/MSEC.2023.3236540.

[356] A. Narayanan and V. Shmatikov. "Robust De-anonymization of Large Sparse Datasets". In: *2008 IEEE Symposium on Security and Privacy*. SP '08. 2008, pp. 111–125. DOI: 10.1109/SP.2008.33.

[357] *National Center for Charitable Statistics Data Archive*. Urban Institute, National Center for Charitable Statistics. 2021. URL: https://nccs-data.urban.org/index.php.

[358] W. Nekoto et al. "Participatory Research for Low-resourced Machine Translation: A Case Study in African Languages". In: *Findings of the Association for Computational Linguistics*. EMNLP '20. 2020, pp. 2144–2160. DOI: 10.18653/v1/2020.findings-emnlp.195.

[359] NetMarketShare. *Market Share Statistics for Internet Technologies*. Feb. 2018. URL: https://netmarketshare.com/.

[360] *NewsGuard*. NewsGuard Technologies. 2021. URL: https://www.newsguardtech.com/.

[361] G. Nicholas and D. Thakur. *Learning to Share. Lessons on Data-Sharing from Beyond Social Media*. Center for Democracy & Technology, Sept. 2022. URL: https://cdt.org/insights/learning-to-share-lessons-on-data-sharing-from-beyond-social-media/.

[362] N. Nikiforakis, L. Invernizzi, A. Kapravelos, S. Van Acker, W. Joosen, C. Kruegel, F. Piessens, and G. Vigna. "You Are What You Include: Large-scale Evaluation of Remote JavaScript Inclusions". In: *19th ACM SIGSAC Conference on Computer and Communications Security*. CCS '12. 2012, pp. 736–747. DOI: 10.1145/2382196.2382274.

[363] Q. Niu, A. Zeng, Y. Fan, and Z. Di. "Robustness of centrality measures against network manipulation". In: *Physica A: Statistical Mechanics and its Applications* 438 (2015), pp. 124–131.

[364] B. A. Nosek, C. R. Ebersole, A. C. DeHaven, and D. T. Mellor. "The preregistration revolution". In: *Proceedings of the National Academy of Sciences* 115.11 (2018), pp. 2600–2606. DOI: 10.1073/pnas.1708274114.

[365] *Online Political Ads: A study of inequality in transparency standards*. Privacy International, Jan. 2021. URL: https://privacyinternational.org/sites/default/files/2021-01/AdsTransparency_TOPUBLISH.pdf.

[366] L. Onwuzurike, E. Mariconti, P. Andriotis, E. D. Cristofaro, G. Ross, and G. Stringhini. "MaMaDroid: Detecting Android Malware by Building Markov Chains of Behavioral Models (Extended Version)". In: *ACM Transactions on Privacy and Security* 22.2 (Apr. 2019). DOI: 10.1145/3313391.

[367] OpenDNS. *OpenDNS System*. URL: https://system.opendns.com/.

[368] *OpenSecrets*. OpenSecrets. 2021. URL: https://www.opensecrets.org/.

[369] *Operation Avalanche: A closer look*. EU publication QP-01-17-801-EN-N. Eurojust, Apr. 2017. DOI: 10.2812/816706. URL: http://www.eurojust.europa.eu/doclibrary/Eurojust-framework/Casework/Operation%20Avalanche%20-%20A%20closer%20look%20(April%202017)/2017-04_Avalanche-Case_EN.pdf.

[370] A.-M. Ortloff, M. Fassl, A. Ponticello, F. Martius, A. Mertens, K. Krombholz, and M. Smith. "Different Researchers, Different Results? Analyzing the Influence of Researcher Experience and Data Type During Qualitative Analysis of an Interview and Survey Study on Security Advice". In: *2023 CHI Conference on Human Factors in Computing Systems*. CHI '23. 2023. DOI: 10.1145/3544548.3580766.

[371] S. Pal. *Sinkholed*. Dec. 3, 2019. URL: https://susam.in/blog/sinkholed/.

[372] K. Papadamou, S. Zannettou, J. Blackburn, E. D. Cristofaro, G. Stringhini, and M. Sirivianos. ""It Is Just a Flu": Assessing the Effect of Watch History on YouTube's Pseudoscientific Video Recommendations". In: *16th International AAAI Conference on Web and Social Media*. ICWSM '22. 2022, pp. 723–734. URL: https://ojs.aaai.org/index.php/ICWSM/article/view/19329.

[373] O. Papakyriakopoulos, A. Gorham, E. Lucherini, M. Kshirsagar, and A. Narayanan. *Facebook's Illusory Promise of Transparency*. Freedom to Tinker. Aug. 5, 2021. URL: https://freedom-to-tinker.com/2021/08/05/facebooks-illusory-promise-of-transparency/.

[374] O. Papakyriakopoulos, C. Tessono, A. Narayanan, and M. Kshirsagar. "How Algorithms Shape the Distribution of Political Advertising: Case Studies of Facebook, Google, and TikTok". In: *2022 AAAI/ACM Conference on AI, Ethics, and Society*. AIES '22. 2022, pp. 532–546. DOI: 10.1145/3514094.3534166.

[375] E. Pariser. *The Filter Bubble: What The Internet Is Hiding From You*. Penguin Press, 2011. ISBN: 9781594203008.

[376] C. Partridge and M. Allman. "Ethical Considerations in Network Measurement Papers". In: *Communications of the ACM* 59.10 (Sept. 2016), pp. 58–64.

[377] E. Pauley and P. McDaniel. "Understanding the Ethical Frameworks of Internet Measurement Studies". In: *2nd International Workshop on Ethics in Computer Security*. EthiCS '23. 2023. DOI: 10.14722/ethics.2023.239547.

[378] A. Pavlo and N. Shi. *Graffiti Networks: A Subversive, Internet-Scale File Sharing Model*. 2011. arXiv: 1101.0350 [cs.NI].

[379] V. Paxson. "Strategies for Sound Internet Measurement". In: *4th ACM SIGCOMM Conference on Internet Measurement*. IMC '04. 2004, pp. 263–271. DOI: 10.1145/1028788.1028824.

[380] F. Pedregosa et al. "Scikit-learn: Machine Learning in Python". In: *Journal of Machine Learning Research* 12 (2011), pp. 2825–2830.

[381] S. Peisert and M. Bishop. "I Am a Scientist, Not a Philosopher!" In: *IEEE Security & Privacy* 5.4 (2007), pp. 48–51. DOI: 10.1109/MSP.2007.84.

[382] P. Peng, L. Yang, L. Song, and G. Wang. "Opening the Blackbox of VirusTotal: Analyzing Online Phishing Scan Engines". In: *2019 Internet Measurement Conference*. IMC '19. 2019, pp. 478–485. DOI: 10.1145/3355369.3355585.

[383] J. Pennekamp, E. Buchholz, M. Dahlmanns, I. Kunze, S. Braun, E. Wagner, M. Brockmann, K. Wehrle, and M. Henze. "Collaboration is not Evil: A Systematic Look at Security Research for Industrial Use". In: *2020 Learning from Authoritative Security Experiment Results Workshop*. LASER '20. 2020. DOI: 10.14722/laser-acsac.2020.23088.

[384] M. Pereira, S. Coleman, B. Yu, M. De Cock, and A. C. A. Nascimento. "Dictionary Extraction and Detection of Algorithmically Generated Domain Names in Passive DNS Traffic". In: *21st International Symposium on Research in Attacks, Intrusions, and Defenses*. RAID '18. 2018, pp. 295–314. DOI: 10.1007/978-3-030-00470-5_14.

[385] N. Petit. "Artificial Intelligence and Automated Law Enforcement: A Review Paper". In: *SSRN Electronic Journal* (2018). DOI: 10.2139/ssrn.3145133. URL: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3145133.

[386] A. Phillips. *Explore Lists in CrowdTangle*. 2021. URL: https://help.crowdtangle.com/en/articles/2450455.

[387] D. Piscitello. *ICANN GDPR and WHOIS Users Survey. A Joint Survey by the Anti-Phishing Working Group (APWG) and the Messaging, Malware and Mobile Anti-Abuse Working Group (M³AAWG)*. Oct. 2018. URL: https://www.m3aawg.org/sites/default/files/m3aawg-apwg-whois-user-survey-report-2018-10.pdf.

[388] D. Plohmann, K. Yakdan, M. Klatt, J. Bader, and E. Gerhards-Padilla. "A Comprehensive Measurement Study of Domain Generating Malware". In: *25th USENIX Security Symposium*. USENIX Security '16. 2016, pp. 263–278.

[389] I. Polakis, F. Maggi, S. Zanero, and A. D. Keromytis. "Security and Privacy Measurements in Social Networks: Experiences and Lessons Learned". In: *3rd International Workshop on Building Analysis Datasets and Gathering Experience Returns for Security*. BADGERS '14. 2014, pp. 18–29. DOI: 10.1109/BADGERS.2014.9.

[390] M. S. Pour, C. Nader, K. Friday, and E. Bou-Harb. "A Comprehensive Survey of Recent Internet Measurement Techniques for Cyber Security". In: *Computers & Security*, 103123 (2023). DOI: 10.1016/j.cose.2023.103123.

[391] "Preliminary Injunction". In: *United States of America v. "flux" a/k/a "ffhost", and "flux2" a/k/a "ffhost2"*. District Court, Western District of Pennsylvania, Dec. 2016. URL: https://www.justice.gov/opa/page/file/917581/download.

[392] *Proposal for a Regulation of the European Parliament and of the Council on a Single Market For Digital Services (Digital Services Act) and amending Directive 2000/31/EC*. COM(2020) 825. Dec. 15, 2020. URL: https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:52020PC0825.

[393] *Proposal for a Regulation of the European Parliament and of the Council on the transparency and targeting of political advertising*. COM(2021) 731. Nov. 25, 2021. URL: https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:52021PC0731.

[394] Quantcast. *Open Internet Ratings Service*. July 5, 2007. URL: https://web.archive.org/web/20070705200342/http://www.quantcast.com/.

[395] M. Z. Rafique, T. Van Goethem, W. Joosen, C. Huygens, and N. Nikiforakis. "It's Free for a Reason: Exploring the Ecosystem of Free Live Streaming Services". In: *23rd Annual Network and Distributed System Security Symposium*. NDSS '16. 2016. DOI: 10.14722/ndss.2016.23030.

[396] L. Rainie and J. Anderson. *Code-Dependent: Pros and Cons of the Algorithm Age. Theme 7: The need grows for algorithmic literacy, transparency and oversight*. Pew Research Center, Feb. 8, 2017. URL: https://www.pewresearch.org/internet/2017/02/08/theme-7-the-need-grows-for-algorithmic-literacy-transparency-and-oversight/.

[397] I. D. Raji, I. E. Kumar, A. Horowitz, and A. Selbst. "The Fallacy of AI Functionality". In: *5th ACM Conference on Fairness, Accountability, and Transparency*. FAccT '22. 2022, pp. 959–972. DOI: 10.1145/3531146.3533158.

[398] Rapid7. *Project Sonar*. URL: https://www.rapid7.com/research/project-sonar/.

[399] E. M. Redmiles, Z. Zhu, S. Kross, D. Kuchhal, T. Dumitras, and M. L. Mazurek. "Asking for a Friend: Evaluating Response Biases in Security User Studies". In: *2018 ACM SIGSAC Conference on Computer and Communications Security*. CCS '18. 2018, pp. 1238–1255. DOI: 10.1145/3243734.3243740.

[400] D. Reidsma, J. van der Ham, and A. Continella. "Operationalizing Cybersecurity Research Ethics Review: From Principles and Guidelines to Practice". In: *2nd International Workshop on Ethics in Computer Security*. EthiCS '23. 2023. DOI: 10.14722/ethics.2023.237352.

[401] F. N. Ribeiro, K. Saha, M. Babaei, L. Henrique, J. Messias, F. Benevenuto, O. Goga, K. P. Gummadi, and E. M. Redmiles. "On Microtargeting Socially Divisive Ads: A Case Study of Russia-Linked Ad Campaigns on Facebook". In: *2nd Conference on Fairness, Accountability, and Transparency*. FAT* '19. 2019, pp. 140–149. DOI: 10.1145/3287560.3287580.

[402] M. H. Ribeiro, R. Ottoni, R. West, V. A. F. Almeida, and W. Meira. "Auditing Radicalization Pathways on YouTube". In: *3rd Conference on Fairness, Accountability, and Transparency*. FAT* '20. 2020, pp. 131–141. DOI: 10.1145/3351095.3372879.

[403] K. Rieck. *Influential Security Papers*. 2022. URL: https://www.sec.tu-bs.de/~konrieck/topnotch/.

[404] A. Rieke and M. Bogen. *Leveling the Platform: Real Transparency for Paid Messages on Facebook*. Upturn, May 2018. URL: https://www.upturn.org/reports/2018/facebook-ads/.

[405] V. Rimmer, D. Preuveneers, M. Juarez, T. Van Goethem, and W. Joosen. "Automated Website Fingerprinting through Deep Learning". In: *25th Annual Network and Distributed System Security Symposium*. NDSS '18. 2018. DOI: 10.14722/ndss.2018.23105.

[406] V. Rimmer, T. Schnitzler, T. Van Goethem, A. Rodríguez Romero, W. Joosen, and K. Kohls. "Trace Oddity: Methodologies for Data-Driven Traffic Analysis on Tor". In: *Proceedings on Privacy Enhancing Technologies* 2022.3 (July 2022), pp. 314–335. DOI: 10.56553/popets-2022-0074.

[407] RIPE NCC Staff. "RIPE Atlas: A Global Internet Measurement Network". In: *The Internet Protocol Journal* 18.3 (Sept. 2015), pp. 2–26. ISSN: 1944-1134. URL: https://ipj.dreamhosters.com/wp-content/uploads/issues/2015/ipj18-3.pdf.

[408] S. Rodota. *Opinion 2/2003 on the application of the data protection principles to the Whois directories*. Article 29 Data Protection Working Party, June 13, 2003. URL: https://ec.europa.eu/justice/article-29/documentation/opinion-recommendation/files/2003/wp76_en.pdf.

[409] G. Rosen, K. Harbath, N. Gleicher, and R. Leathern. *Helping to Protect the 2020 US Elections*. Facebook. Oct. 21, 2019. URL: https://about.fb.com/news/2019/10/update-on-election-integrity-efforts/.

[410] E. Rosenberg. "Facebook blocked many gay-themed ads as part of its new advertising policy, angering LGBT groups". In: *The Washington Post* (Oct. 3, 2018). URL: https://www.washingtonpost.com/technology/2018/10/03/facebook-blocked-many-gay-themed-ads-part-its-new-advertising-policy-angering-lgbt-groups/.

[411] M. Rosenberg. "Ad Tool Facebook Built to Fight Disinformation Doesn't Work as Advertised". In: *The New York Times* (July 25, 2019). URL: https://www.nytimes.com/2019/07/25/technology/facebook-ad-library.html.

[412] C. Rossow, C. J. Dietrich, C. Grier, C. Kreibich, V. Paxson, N. Pohlmann, H. Bos, and M. v. Steen. "Prudent Practices for Designing Malware Experiments: Status Quo and Outlook". In: *2012 IEEE Symposium on Security and Privacy*. SP '12. 2012, pp. 65–79. DOI: 10.1109/SP.2012.14.

[413] S. Roth, S. Calzavara, M. Wilhelm, A. Rabitti, and B. Stock. "The Security Lottery: Measuring Client-Side Web Security Inconsistencies". In: *31st USENIX Security Symposium*. USENIX Security '22. 2022, pp. 2047–2064.

[414] K. Ruth, A. Fass, J. Azose, M. Pearson, E. Thomas, C. Sadowski, and Z. Durumeric. "A World Wide View of Browsing the World Wide Web". In: *22nd ACM Internet Measurement Conference*. IMC '22. 2022. DOI: 10.1145/3517745.3561418.

[415] K. Ruth, D. Kumar, B. Wang, L. Valenta, and Z. Durumeric. "Toppling Top Lists: Evaluating the Accuracy of Popular Website Lists". In: *22nd ACM Internet Measurement Conference*. IMC '22. 2022. DOI: 10.1145/3517745.3561444.

[416] W. Rweyemamu, T. Lauinger, C. Wilson, W. Robertson, and E. Kirda. "Clustering and the Weekend Effect: Recommendations for the Use of Top Domain Lists in Security Research". In: *20th International Conference on Passive and Active Measurement*. PAM '19. 2019, pp. 161–177. DOI: 10.1007/978-3-030-15986-3_11.

[417] W. Rweyemamu, T. Lauinger, C. Wilson, W. Robertson, and E. Kirda. "Getting Under Alexa's Umbrella: Infiltration Attacks Against Internet Top Domain Lists". In: *22nd International Conference on Information Security*. ISC '19. 2019, pp. 255–276.

[418] *Safe Browsing*. Google Inc. URL: https://safebrowsing.google.com/.

[419] A. San, J. Bakus, C. Lockard, D. Ciemiewicz, Y. Ji, S. Atluri, K. Small, and H. Elfardy. *PLAtE: A Large-scale Dataset for List Page Web Extraction*. 2022. DOI: 10.48550/arxiv.2205.12386. arXiv: 2205.12386.

[420] C. Sandvig, K. Hamilton, K. Karahalios, and C. Langbort. "Auditing Algorithms: Research Methods for Detecting Discrimination on Internet Platforms". In: *Data and discrimination: Converting Critical Concerns into Productive Inquiry*. 2014. URL: http://www-personal.umich.edu/~csandvig/research/Auditing%20Algorithms%20--%20Sandvig%20--%20ICA%202014%20Data%20and%20Discrimination%20Preconference.pdf.

[421] A. Sankin and S. Mattu. "The High Privacy Cost of a "Free" Website". In: *The Markup* (Sept. 22, 2020). URL: https://themarkup.org/blacklight/2020/09/22/blacklight-tracking-advertisers-digital-privacy-sensitive-websites.

[422] P. Sapiezynski, A. Ghosh, L. Kaplan, A. Rieke, and A. Mislove. "Algorithms That "Don't See Color": Measuring Biases in Lookalike and Special Ad Audiences". In: *2022 AAAI/ACM Conference on AI, Ethics, and Society*. AIES '22. 2022, pp. 609–616. DOI: 10.1145/3514094.3534135.

[423] D. Saucez and L. Iannone. "Thoughts and Recommendations from the ACM SIGCOMM 2017 Reproducibility Workshop". In: *ACM SIGCOMM Computer Communication Review* 48.1 (Apr. 2018), pp. 70–74. DOI: 10.1145/3211852.3211863.

[424] S. Schechter. *Common Pitfalls in Writing about Security and Privacy Human Subjects Experiments, and How to Avoid Them*. MSR-TR-2013-5. Jan. 2013. URL: https://www.microsoft.com/en-us/research/publication/common-pitfalls-in-writing-about-security-and-privacy-human-subjects-experiments-and-how-to-avoid-them/.

[425] J. Scheck, N. Purnell, and J. Horwitz. "Facebook Employees Flag Drug Cartels and Human Traffickers. The Company's Response Is Weak, Documents Show." In: *The Wall Street Journal* (Sept. 16, 2021). URL: https://www.wsj.com/articles/facebook-drug-cartels-human-traffickers-response-is-weak-documents-11631812953.

[426] Q. Scheitle, O. Gasser, T. Nolte, J. Amann, L. Brent, G. Carle, R. Holz, T. C. Schmidt, and M. Wählisch. "The Rise of Certificate Transparency and Its Implications on the Internet Ecosystem". In: *18th Internet Measurement Conference*. IMC '18. 2018, pp. 343–349. DOI: 10.1145/3278532.3278562.

[427] Q. Scheitle, O. Hohlfeld, J. Gamba, J. Jelten, T. Zimmermann, S. D. Strowes, and N. Vallina-Rodriguez. "A Long Way to the Top: Significance, Structure, and Stability of Internet Top Lists". In: *18th Internet Measurement Conference*. IMC '18. 2018, pp. 478–493. DOI: 10.1145/3278532.3278574.

[428] Q. Scheitle, M. Wählisch, O. Gasser, T. C. Schmidt, and G. Carle. "Towards an Ecosystem for Reproducible Research in Computer Networking". In: *Reproducibility Workshop*. Reproducibility '17. 2017, pp. 5–8. DOI: 10.1145/3097766.3097768.

[429] S. Schiavoni, F. Maggi, L. Cavallaro, and S. Zanero. "Phoenix: DGA-Based Botnet Tracking and Intelligence". In: *11th International Conference on Detection of Intrusions and Malware, and Vulnerability Assessment*. DIMVA '14. 2014, pp. 192–211. DOI: 10.1007/978-3-319-08509-8_11.

[430] M. Schmidt. *Expired Domains*. Aug. 2018. URL: https://www.expireddomains.net.

[431] H. Schulzrinne. "Double-blind reviewing". In: *ACM SIGCOMM Computer Communication Review* 39.2 (Mar. 2009), pp. 56–59. DOI: 10.1145/1517480.1517492.

[432] S. Schüppen, D. Teubert, P. Herrmann, and U. Meyer. "FANCI : Feature-based Automated NXDomain Classification and Intelligence". In: *27th USENIX Security Symposium*. USENIX Security '18. 2018, pp. 1165–1181.

[433] D. Schwarz. *Bedep's DGA: Trading Foreign Exchange for Malware Domains*. Arbor Networks. Apr. 21, 2015. URL: https://web.archive.org/web/20160114122355/https://asert.arbornetworks.com/bedeps-dga-trading-foreign-exchange-for-malware-domains/.

[434] M. Scott. "Political ads on Facebook disappear ahead of UK election". In: *Politico* (Dec. 10, 2019). URL: https://www.politico.com/news/2019/12/10/political-ads-on-facebook-disappear-ahead-of-uk-election-081376.

[435] M. Scott and Z. Montellaro. "Scores of political groups sidestepped Facebook's ad ban". In: *Politico* (Mar. 4, 2021). URL: https://www.politico.com/news/2021/03/04/politcal-groups-facebook-ad-ban-473698.

[436] N. B. Shah. "Challenges, Experiments, and Computational Solutions in Peer Review". In: *Communications of the ACM* 65.6 (May 2022), pp. 76–87. DOI: 10.1145/3528086.

[437] R. Shirazi. "Botnet Takedown Initiatives: A Taxonomy and Performance Model". In: *Technology Innovation Management Review* 5.1 (Jan. 2015), pp. 15–20. ISSN: 1927-0321. DOI: 10.22215/timreview/862.

[438] M. Silva and F. Benevenuto. "COVID-19 Ads as Political Weapon". In: *36th ACM/SIGAPP Symposium on Applied Computing*. SAC '21. 2021, pp. 1705–1710. DOI: 10.1145/3412841.3442043.

[439] M. Silva, L. Santos de Oliveira, A. Andreou, P. O. Vaz de Melo, O. Goga, and F. Benevenuto. "Facebook Ads Monitor: An Independent Auditing System for Political Ads on Facebook". In: *The Web Conference 2020*. WWW '20. 2020, pp. 224–234. DOI: 10.1145/3366423.3380109.

[440] C. Silverman and R. Mac. "Facebook Promised To Label Political Ads, But Ads For Biden, The Daily Wire, And Interest Groups Are Slipping Through". In: *BuzzFeed News* (Oct. 22, 2020). URL: https://www.buzzfeednews.com/article/craigsilverman/facebook-biden-election-ads.

[441] J. Simko, M. Tomlein, B. Pecher, R. Moro, I. Srba, E. Stefancova, A. Hrckova, M. Kompan, J. Podrouzek, and M. Bielikova. "Towards Continuous Automatic Audits of Social Media Adaptive Behavior and Its Role in Misinformation Spreading". In: *Adjunct Proceedings of the 29th ACM Conference on User Modeling, Adaptation and Personalization*. UMAP '21 Adjunct. 2021, pp. 411–414. DOI: 10.1145/3450614.3463353.

[442] S. Simpson. *For sites that are not Quantified, it says "data is estimated." What does this mean?* Jan. 26, 2018. URL: https://quantcast.zendesk.com/hc/en-us/articles/115013961667.

[443] S. Simpson. *What is the date range for the traffic numbers on your site?* Jan. 26, 2018. URL: https://quantcast.zendesk.com/hc/en-us/articles/115013961687.

[444] S. Sinha, M. Bailey, and F. Jahanian. "Shades of grey: On the effectiveness of reputation-based "blacklists"". In: *3rd International Conference on Malicious and Unwanted Software*. MALWARE '08. 2008, pp. 57–64. DOI: 10.1109/MALWARE.2008.4690858.

[445] R. Sion. *Democracy in Peer Reviewing*. 2011. URL: https://zxr.io/theseriousacademic/.

[446] *Sistema de Consulta a CNPJs*. Tribunal Superior Eleitoral. 2020. URL: https://spce-cnpj.tse.jus.br/spce2016.cnpj/internet/#/eleicoes.

[447] R. Sivaguru, C. Choudhary, B. Yu, V. Tymchenko, A. Nascimento, and M. De Cock. "An Evaluation of DGA Classifiers". In: *2018 IEEE International Conference on Big Data*. Big Data '18. 2018, pp. 5058–5067. DOI: 10.1109/BigData.2018.8621875.

[448] G. Sloane. "Facebook repeatedly blocked ads showing wheelchair, says disabilities apparel retailer". In: *Ad Age* (Dec. 21, 2020). URL: https://adage.com/article/digital/facebook-repeatedly-blocked-ads-showing-wheelchair-says-disabilities-apparel-retailer/2302536.

[449] S. Smalley. "Meta won't comment on its plans to abandon CrowdTangle". In: *Poynter* (Aug. 18, 2022). URL: https://www.poynter.org/reporting-editing/2022/meta-wont-comment-on-its-plans-to-abandon-crowdtangle/.

[450] S. Smith. "Facebook denies Houston's ads promoting fair housing over race, religion references". In: *Houston Chronicle* (May 15, 2019). URL: https://www.houstonchronicle.com/news/houston-texas/houston/article/Facebook-denies-Houston-s-ads-promoting-fair-13847765.php.

[451] *Social media companies are failing to provide adequate advertising transparency to users globally*. Privacy International, Oct. 3, 2019. URL: https://privacyinternational.org/sites/default/files/2019-10/cop-2019_0.pdf.

[452] R. Sommer and V. Paxson. "Outside the Closed World: On Using Machine Learning for Network Intrusion Detection". In: *2010 IEEE Symposium on Security and Privacy*. SP '10. 2010, pp. 305–316. DOI: 10.1109/SP.2010.25.

[453] A. Soneji, F. B. Kokulu, C. Rubio-Medrano, T. Bao, R. Wang, Y. Shoshitaishvili, and A. Doupé. ""Flawed, but like democracy we don't have a better system": The Experts' Insights on the Peer Review Process of Evaluating Security Papers". In: *2022 IEEE Symposium on Security and Privacy*. SP '22. 2022, pp. 1845–1862. DOI: 10.1109/SP46214.2022.9833581.

[454] M. Sonntag. "DNS Traffic of a Tor Exit Node - An Analysis". In: *Security, Privacy, and Anonymity in Computation, Communication, and Storage*. SpaCCS '18. 2018, pp. 33–45. DOI: 10.1007/978-3-030-05345-1_3.

[455] K. Soska and N. Christin. "Automatically Detecting Vulnerable Websites Before They Turn Malicious". In: *23rd USENIX Security Symposium*. USENIX Security '14. 2014, pp. 625–640.

[456] V. Sosnovik and O. Goga. "Understanding the Complexity of Detecting Political Ads". In: *The Web Conference 2021*. WWW '21. 2021, pp. 2002–2013. DOI: 10.1145/3442381.3450049.

[457] A. Sperotto, O. van der Toorn, and R. van Rijswijk-Deij. "TIDE: Threat Identification Using Active DNS Measurements". In: *Proceedings of the SIGCOMM Posters and Demos*. SIGCOMM Posters and Demos '17. 2017, pp. 65–67. DOI: 10.1145/3123878.3131988.

[458] L. Spinelli and M. Crovella. "How YouTube Leads Privacy-Seeking Users Away from Reliable Information". In: *Adjunct Publication of the 28th ACM Conference on User Modeling, Adaptation and Personalization*. UMAP '20 Adjunct. 2020, pp. 244–251. DOI: 10.1145/3386392.3399566.

[459] J. Spooren, D. Preuveneers, L. Desmet, P. Janssen, and W. Joosen. "Detection of Algorithmically Generated Domain Names Used by Botnets: A Dual Arms Race". In: *34th ACM/SIGAPP Symposium on Applied Computing*. SAC '19. 2019, pp. 1916–1923. DOI: 10.1145/3297280.3297467.

[460] J. Spooren, T. Vissers, P. Janssen, W. Joosen, and L. Desmet. "Premadoma: An Operational Solution for DNS Registries to Prevent Malicious Domain Registrations". In: *35th Annual Computer Security Applications Conference*. ACSAC '19. 2019, pp. 557–567. DOI: 10.1145/3359789.3359836.

[461] M. Stampar. *Email addresses used in WHOIS registrations of sinkholed malicious/malware domains*. Oct. 2018. URL: https://gist.github.com/stamparm/9726d93fd0048aee6c54ec88a8e85bfc.

[462] M. Stampar et al. *maltrail: Malicious traffic detection system*. 2019. URL: https://github.com/stamparm/maltrail.

[463] M. A. Stelzner. *2020 Social Media Marketing Industry Report*. Social Media Examiners, May 2020. URL: https://alltomalmhult.se/wp-content/uploads/2020/05/IndustryReport-2020.pdf.

[464] M. Stevanovic, J. M. Pedersen, A. D'Alconzo, S. Ruehrup, and A. Berger. "On the ground truth problem of malicious DNS traffic analysis". In: *Computers & Security* 55 (2015), pp. 142–158. DOI: 10.1016/j.cose.2015.09.004.

[465] S. Sugrim, C. Liu, M. McLean, and J. Lindqvist. "Robust Performance Metrics for Authentication Systems". In: *2019 Network and Distributed System Security Symposium*. 2019. DOI: 10.14722/ndss.2019.23351.

[466] L. Sweeney. "Discrimination in Online Ad Delivery". In: *Communications of the ACM* 56.5 (May 2013), pp. 44–54. DOI: 10.1145/2447976.2447990.

[467] J. Szurdi, M. Luo, B. Kondracki, N. Nikiforakis, and N. Christin. "Where are you taking me? Understanding Abusive Traffic Distribution Systems". In: *The Web Conference 2021*. WWW '21. 2021, pp. 3613–3624. DOI: 10.1145/3442381.3450071.

[468] D. Tahir. "Facebook political ad ban blocks pro-vaccine messages". In: *Politico* (Feb. 21, 2021). URL: https://www.politico.com/news/2021/02/21/facebook-ad-pro-vaccine-ban-470304.

[469] S. Tajalizadehkhoob, M. Korczyński, A. Noroozian, C. Gañán, and M. van Eeten. "Apples, oranges and hosting providers: Heterogeneity and security in the hosting market". In: *2016 IEEE/IFIP Network Operations and Management Symposium*. NOMS '16. 2016, pp. 289–297. DOI: 10.1109/NOMS.2016.7502824.

[470] M. Taylor. *TLS 1.0 and 1.1 Removal Update*. Mozilla Hacks. May 15, 2019. URL: https://hacks.mozilla.org/2019/05/tls-1-0-and-1-1-removal-update/.

[471] L. Teixeira da Rocha. *InfoRanks: Infoblox Ranking Service. Statistical inference for defining internet ranks*. Infoblox, 2021. URL: https://www.infoblox.com/wp-content/uploads/infoblox-whitepaper-inforanks-infoblox-ranking-service.pdf.

[472] W. Thayer. *Delaying Further Symantec TLS Certificate Distrust*. Mozilla Foundation. Oct. 2018. URL: https://blog.mozilla.org/security/2018/10/10/delaying-further-symantec-tls-certificate-distrust/.

[473] The Tor Project. *Number or relays with relay flags assigned*. Apr. 26, 2018. URL: https://metrics.torproject.org/relayflags.html?start=2017-01-01&end=2018-04-26&flag=Exit.

[474] J. Tierney. "Do You Suffer From Decision Fatigue?" In: *The New York Times Magazine* (Aug. 17, 2011). URL: https://www.nytimes.com/2011/08/21/magazine/do-you-suffer-from-decision-fatigue.html.

[475] M. Tomlein, B. Pecher, J. Simko, I. Srba, R. Moro, E. Stefancova, M. Kompan, A. Hrckova, J. Podrouzek, and M. Bielikova. "An Audit of Misinformation Filter Bubbles on YouTube: Bubble Bursting and Recent Behavior Changes". In: *15th ACM Conference on Recommender Systems*. RecSys '21. 2021, pp. 1–11. DOI: 10.1145/3460231.3474241.

[476] L. Trahan. *Trahan Leads Introduction of Social Media DATA Transparency Legislation*. May 20, 2021. URL: https://trahan.house.gov/news/documentsingle.aspx?DocumentID=2112.

[477] *Tranco*. SEKOIA.IO. 2022. URL: https://docs.sekoia.io/tip/features/automate/library/tranco/.

[478] R. Tromble. *A Paucity of Data: The Digital Platforms' Responses to Pillar 5 of the Code of Practice on Disinformation*. The George Washington University, May 4, 2020. URL: https://iddp.gwu.edu/paucity-data.

[479] *URLhaus API*. abuse.ch. 2022. URL: https://urlhaus.abuse.ch/api/.

[480] *US 2020 Elections Ad Spending Tracker*. Facebook. Jan. 23, 2021. URL: https://about.fb.com/wp-content/uploads/2021/01/US-2020-Elections-Ad-Spending-Tracker.zip.

[481] *US Ad Restriction Period*. Facebook. Oct. 2020. URL: https://www.facebook.com/gms_hub/share/facebook-us-ad-restriction-period-overview.pdf.

[482] *USENIX Security '22 Call for Artifacts*. USENIX Association. July 27, 2022. URL: https://www.usenix.org/conference/usenixsecurity22/call-for-artifacts.

[483] P. Vadrevu and R. Perdisci. "What You See is NOT What You Get: Discovering and Tracking Social Engineering Attack Campaigns". In: *19th Internet Measurement Conference*. IMC '19. 2019, pp. 308–321. DOI: 10.1145/3355369.3355600.

[484] J. Valentino-DeVries. "I Approved This Facebook Message – But You Don't Know That". In: *ProPublica* (Feb. 13, 2018). URL: https://www.propublica.org/article/i-approved-this-facebook-message-but-you-dont-know-that.

[485] J. Valentino-DeVries. *New Computer Bug Exposes Broad Security Flaws*. The Wall Street Journal. May 19, 2015. URL: https://www.wsj.com/articles/new-computer-bug-exposes-broad-security-flaws-1432076565.

[486] P. Vallina, V. Le Pochat, Á. Feal, M. Paraschiv, J. Gamba, T. Burke, O. Hohlfeld, J. Tapiador, and N. Vallina-Rodriguez. "Mis-shapes, Mistakes, Misfits: An Analysis of Domain Classification Services". In: *20th Internet Measurement Conference*. IMC '20. 2020, pp. 598–618. DOI: 10.1145/3419394.3423660.

[487] T. Van Goethem, V. Le Pochat, and W. Joosen. "Mobile Friendly or Attacker Friendly?: A Large-scale Security Evaluation of Mobile-first Websites". In: *2019 ACM Asia Conference on Computer and Communications Security*. AsiaCCS '19. 2019, pp. 206–213. DOI: 10.1145/3321705.3329855.

[488] T. Van hamme, G. Garofalo, S. Joos, D. Preuveneers, and W. Joosen. "AI for Biometric Authentication Systems". In: *Security and Artificial Intelligence: A Crossdisciplinary Approach*. 2022, pp. 156–180. ISBN: 978-3-030-98795-4. DOI: 10.1007/978-3-030-98795-4_8.

[489] R. van Rijswijk-Deij, M. Jonker, A. Sperotto, and A. Pras. "A High-Performance, Scalable Infrastructure for Large-Scale Active DNS Measurements". In: *IEEE Journal on Selected Areas in Communications* 34.6 (June 2016), pp. 1877–1888. DOI: 10.1109/JSAC.2016.2558918.

[490] B. VanderSloot, J. Amann, M. Bernhard, Z. Durumeric, M. Bailey, and J. A. Halderman. "Towards a Complete View of the Certificate Ecosystem". In: *16th Internet Measurement Conference*. IMC '16. 2016, pp. 543–549. DOI: 10.1145/2987443.2987462.

[491] M. Y. Vardi. "Academic Rankings Considered Harmful!" In: *Communications of the ACM* 59.9 (Aug. 2016), p. 5. DOI: 10.1145/2980760.

[492] M. Y. Vardi. "Conferences vs. Journals in Computing Research". In: *Communications of the ACM* 52.5 (May 2009), p. 5. DOI: 10.1145/1506409.1506410.

[493] G. Venkatadri, A. Andreou, Y. Liu, A. Mislove, K. P. Gummadi, P. Loiseau, and O. Goga. "Privacy Risks with Facebook's PII-Based Targeting: Auditing a Data Broker's Advertising Interface". In: *2018 IEEE Symposium on Security and Privacy*. SP '18. 2018, pp. 89–107. DOI: 10.1109/SP.2018.00014.

[494] Verisign, ed. *The Domain Name Industry Brief* 19 (2 June 2022). URL: https://www.verisign.com/assets/domain-name-report-Q12022.pdf.

[495] T. Verma and S. Singanamalla. *Improving DNS Privacy with Oblivious DoH in 1.1.1.1*. Cloudflare. Dec. 8, 2020. URL: https://blog.cloudflare.com/oblivious-dns/.

[496] J. Vincent. "Amazon is retiring Alexa – no, not that one". In: *The Verge* (Dec. 9, 2021). URL: https://www.theverge.com/2021/12/9/22825744/amazon-retiring-alexa-web-ranking-sevice.

[497] *Virtual Insanity? The need to guarantee transparency in digital political advertising*. European Partnership for Democracy, Mar. 2020. URL: https://epd.eu/virtual-insanity/.

[498] T. Vissers, T. Barron, T. Van Goethem, W. Joosen, and N. Nikiforakis. "The Wolf of Name Street: Hijacking Domains Through Their Nameservers". In: *2017 ACM SIGSAC Conference on Computer and Communications Security*. 2017, pp. 957–970. DOI: 10.1145/3133956.3133988.

[499] T. Vissers, W. Joosen, and N. Nikiforakis. "Parking Sensors: Analyzing and Detecting Parked Domains". In: *22nd Annual Network and Distributed System Security Symposium*. 2015. DOI: 10.14722/ndss.2015.23053.

[500] D. Vrandečić and M. Krötzsch. "Wikidata: A Free Collaborative Knowledgebase". In: *Communications of the ACM* 57.10 (Sept. 2014), pp. 78–85. DOI: 10.1145/2629489.

[501] C. de Vreese and R. Tromble. "The Data Abyss: How Lack of Data Access Leaves Research and Society in the Dark". In: *Political Communication* (May 2023), pp. 1–5. DOI: 10.1080/10584609.2023.2207488.

[502] S. Vrhovec, L. Caviglione, and S. Wendzel. "Crème de La Crème: Lessons from Papers in Security Publications". In: *16th International Conference on Availability, Reliability and Security*. ARES '21. 2021, 92. DOI: 10.1145/3465481.3470027.

[503] S. Wachter, B. Mittelstadt, and L. Floridi. "Why a Right to Explanation of Automated Decision-Making Does Not Exist in the General Data Protection Regulation". In: *International Data Privacy Law* 7.2 (May 2017), pp. 76–99. DOI: 10.1093/idpl/ipx005.

[504] K. Wagner and N. Nix. "Facebook Disables Accounts Tied to NYU Research Project". In: *Bloomberg* (Aug. 3, 2021). URL: https://www.bloomberg.com/news/articles/2021-08-03/facebook-disables-accounts-tied-to-nyu-research-project.

[505] R. Wainwright and F. J. Cilluffo. *Responding to Cybercrime at Scale: Operation Avalanche - A Case Study*. Issue Brief 2017-03. Europol & Center for Cyber and Homeland Security, The George Washington University, Mar. 2017. URL: https://cchs.gwu.edu/sites/g/files/zaxdzs2371/f/Responding%20to%20Cybercrime%20at%20Scale%20FINAL.pdf.

[506] G. Wan, L. Izhikevich, D. Adrian, K. Yoshioka, R. Holz, C. Rossow, and Z. Durumeric. "On the Origin of Scanning: The Impact of Location on Internet-Wide Scans". In: *20th Internet Measurement Conference*. IMC '20. 2020, pp. 662–679. DOI: 10.1145/3419394.3424214.

[507] D. Warburton. *The 2021 TLS Telemetry Report*. F5 Labs. Oct. 20, 2021. URL: https://www.f5.com/labs/articles/threat-intelligence/the-2021-tls-telemetry-report.

[508] H. J. Watson and C. Nations. "Addressing the Growing Need for Algorithmic Transparency". In: *Communications of the Association for Information Systems* (2019), pp. 488–510. DOI: 10.17705/1cais.04526.

[509] N. Watzman. *The political ads Facebook won't show you*. Cybersecurity for Democracy. May 12, 2021. URL: https://medium.com/cybersecurity-for-democracy/e0d6181bca25.

[510] *Wayback Machine APIs*. The Internet Archive. Sept. 24, 2013. URL: https://archive.org/help/wayback_api.php.

[511] *Web Almanac*. HTTP Archive. 2022. URL: https://almanac.httparchive.org/.

[512] M. Webb and B. John. *We need to know more about political ads. But can transparency be a trap?* First Draft Footnotes. Mar. 25, 2021. URL: https://medium.com/1st-draft/542df2a52f21.

[513] W. Webber, A. Moffat, and J. Zobel. "A similarity measure for indefinite rankings". In: *ACM Transactions on Information Systems* 28.4, 38 (Nov. 2010), p. 20. DOI: 10.1145/1852102.1852106.

[514] S. Wendzel, C. Lévy-Bencheton, and L. Caviglione. "Not All Areas Are Equal: Analysis of Citations in Information Security Research". In: *Scientometrics* 122.1 (Jan. 2020), pp. 267–286. DOI: 10.1007/s11192-019-03279-6.

[515] *What is the Facebook Ad Library and how do I search it?* Facebook Help Center. 2021. URL: https://www.facebook.com/help/259468828226154.

[516] A. Wheeler. "Facebook under fire after ads for anti-HIV drug PrEP deemed political". In: *The Guardian* (Oct. 31, 2019). URL: https://www.theguardian.com/technology/2019/oct/31/facebook-prep-ads-instagram-political.

[517] *WHOIS and Data Protection*. ICANN Governmental Advisory Committee. Oct. 7, 2021. URL: https://gac.icann.org/activity/whois-and-data-protection.

[518] T. Wicinski. *DNS Privacy Considerations*. RFC 9076. July 2021. DOI: 10.17487/RFC9076. URL: https://www.rfc-editor.org/info/rfc9076.

[519] G. Widmer and M. Kubat. "Learning in the Presence of Concept Drift and Hidden Contexts". In: *Machine Learning* 23.1 (Apr. 1996), pp. 69–101. DOI: 10.1023/A:1018046501280.

[520] M. D. Wilkinson et al. "The FAIR Guiding Principles for scientific data management and stewardship". In: *Scientific Data* 3.1 (Mar. 2016). DOI: 10.1038/sdata.2016.18.

[521] C. Wilson, A. Ghosh, S. Jiang, A. Mislove, L. Baker, J. Szary, K. Trindel, and F. Polli. "Building and Auditing Fair Algorithms: A Case Study in Candidate Screening". In: *4th ACM Conference on Fairness, Accountability, and Transparency*. FAccT '21. 2021, pp. 666–677. DOI: 10.1145/3442188.3445928.

[522] J. Woodbridge, H. S. Anderson, A. Ahuja, and D. Grant. *Predicting Domain Generation Algorithms with Long Short-Term Memory Networks*. Nov. 2016. arXiv: 1611.00791 [cs.CR].

[523] Q. Xie, S. Tang, X. Zheng, Q. Lin, B. Liu, H. Duan, and F. Li. "Building an Open, Robust, and Stable Voting-Based Domain Top List". In: *31st USENIX Security Symposium*. USENIX Security '22. 2022, pp. 625–642.

[524] W. Xu, K. Sanders, and Y. Zhang. "We know it before you do: Predicting malicious domains". In: *Virus Bulletin Conference*. Sept. 2014, pp. 73–77.

[525] S. Yadav and A. L. N. Reddy. "Winning with DNS Failures: Strategies for Faster Botnet Detection". In: *7th International ICST Conference on Security and Privacy in Communication Networks*. SecureComm '11. 2011, pp. 446–459. DOI: 10.1007/978-3-642-31909-9_26.

[526] S. Yadav, A. K. K. Reddy, A. N. Reddy, and S. Ranjan. "Detecting Algorithmically Generated Domain-Flux Attacks With DNS Traffic Analysis". In: *IEEE/ACM Transactions on Networking* 20.5 (Oct. 2012), pp. 1663–1677. DOI: 10.1109/TNET.2012.2184552.

[527] S. Yadav, A. K. K. Reddy, A. N. Reddy, and S. Ranjan. "Detecting Algorithmically Generated Malicious Domain Names". In: *10th ACM SIGCOMM Conference on Internet Measurement*. IMC '10. 2010, pp. 48–61. DOI: 10.1145/1879141.1879148.

[528] J. Yesbeck. *Your Top Questions About Alexa Data and Ranks, Answered*. Oct. 29, 2014. URL: https://blog.alexa.com/top-questions-about-alexa-answered/.

[529] D. Zeber, S. Bird, C. Oliveira, W. Rudametkin, I. Segall, F. Wollsén, and M. Lopatka. "The Representativeness of Automated Web Crawls as a Surrogate for Human Browsing". In: *The Web Conference 2020*. WWW '20. 2020, pp. 167–178. DOI: 10.1145/3366423.3380104.

[530] E. Zeng, T. Kohno, and F. Roesner. "What Makes a "Bad" Ad? User Perceptions of Problematic Online Advertising". In: *2021 CHI Conference on Human Factors in Computing Systems*. CHI '21. 2021. DOI: 10.1145/3411764.3445459.

[531] E. Zeng, M. Wei, T. Gregersen, T. Kohno, and F. Roesner. "Polls, Clickbait, and Commemorative $2 Bills: Problematic Political Advertising on News and Media Websites around the 2020 U.S. Elections". In: *21st ACM Internet Measurement Conference*. IMC '21. 2021, pp. 507–525. DOI: 10.1145/3487552.3487850.

[532] X. Zhang, X. Wang, X. Bai, Y. Zhang, and X. Wang. "OS-level Side Channels without Procfs: Exploring Cross-App Information Leakage on iOS". In: *25th Annual Network and Distributed System Security Symposium*. NDSS '18. 2018. DOI: 10.14722/ndss.2018.23260.

[533] Y. Zhang, M. Liu, M. Zhang, C. Lu, and H. Duan. "Ethics in Security Research: Visions, Reality, and Paths Forward". In: *2022 IEEE European Symposium on Security and Privacy Workshops*. EuroS&PW '22. 2022, pp. 538–545. DOI: 10.1109/EuroSPW55150.2022.00064.

[534] B. Z. H. Zhao, M. Ikram, H. J. Asghar, M. A. Kaafar, A. Chaabane, and K. Thilakarathna. "A Decade of Mal-Activity Reporting: A Retrospective Analysis of Internet Malicious Activity Blacklists". In: *14th ACM Asia Conference on Computer and Communications Security*. ASIACCS '19. 2019, pp. 193–205. DOI: 10.1145/3321705.3329834.

[535] Y. Zhauniarovich, I. Khalil, T. Yu, and M. Dacier. "A Survey on Malicious Domains Detection Through DNS Data Analysis". In: *ACM Computing Surveys* 51.4, 67 (July 2018). DOI: 10.1145/3191329.

[536] M. Zheng, H. Robbins, Z. Chai, P. Thapa, and T. Moore. "Cybersecurity Research Datasets: Taxonomy and Empirical Analysis". In: *11th USENIX Workshop on Cyber Security Experimentation and Test*. CSET '18. 2018. URL: https://www.usenix.org/conference/cset18/presentation/zheng.

[537] J. Zhou. *Top Cyber Security Conferences Ranking*. 2022. URL: http://jianying.space/conference-ranking.html.

[538] S. Zhu, J. Shi, L. Yang, B. Qin, Z. Zhang, L. Song, and G. Wang. "Measuring and Modeling the Label Dynamics of Online Anti-Malware Engines". In: *29th USENIX Security Symposium*. USENIX Security '20. 2020, pp. 2361–2378. URL: https://www.usenix.org/conference/usenixsecurity20/presentation/zhu.

[539] N. Zilberman and A. W. Moore. "Thoughts about Artifact Badging". In: *ACM SIGCOMM Computer Communication Review* 50.2 (May 2020), pp. 60–63. DOI: 10.1145/3402413.3402422.

[540] S. Zimmeck, J. S. Li, H. Kim, S. M. Bellovin, and T. Jebara. "A Privacy Analysis of Cross-device Tracking". In: *26th USENIX Security Symposium*. USENIX Security '17. 2017, pp. 1391–1408.

List of publications

Publications in peer-reviewed conference proceedings:

1. Victor Le Pochat, Tom Van Goethem, Wouter Joosen. "Idea: Visual Analytics for Web Security". 10th International Symposium on Engineering Secure Software and Systems (ESSoS 2018). DOI: 10.1007/978-3-319-94496-8_10

2. Victor Le Pochat, Tom Van Goethem, Samaneh Tajalizadehkhoob, Maciej Korczyński, Wouter Joosen. "Tranco: A Research-Oriented Top Sites Ranking Hardened Against Manipulation". 26th Annual Network and Distributed System Security Symposium (NDSS 2019). DOI: 10.14722/ndss.2019.23386.
ACSAC 2022 Cybersecurity Artifacts Competition Impactful Dataset Award

3. Victor Le Pochat, Tom Van Goethem, Wouter Joosen. "Funny Accents: Exploring Genuine Interest in Internationalized Domain Names". 20th Passive and Active Measurement Conference (PAM 2019). DOI: 10.1007/978-3-030-15986-3_12

4. Tom Van Goethem*, Victor Le Pochat*, Wouter Joosen. *(* Joint first author.)* "Mobile Friendly or Attacker Friendly? A Large-scale Security Evaluation of Mobile-first Websites". 2019 ACM Asia Conference on Computer and Communications Security (AsiaCCS 2019). DOI: 10.1145/3321705.3329855

5. Victor Le Pochat, Tim Van hamme, Sourena Maroofi, Tom Van Goethem, Davy Preuveneers, Andrzej Duda, Wouter Joosen, Maciej Korczyński. "A Practical Approach for Taking Down Avalanche Botnets Under Real-World Constraints". 27th Annual Network and Distributed System Security Symposium (NDSS 2020). DOI: 10.14722/ndss.2020.24161

6. Pelayo Vallina, Victor Le Pochat, Álvaro Feal, Marius Paraschiv, Julien Gamba, Tim Burke, Oliver Hohlfeld, Juan Tapiador, Narseo Vallina-Rodriguez. "Mis-shapes, Mistakes, Misfits: An Analysis of Domain Classification Services". 2020 Internet Measurement Conference (IMC 2020). DOI: 10.1145/3419394.3423660

7. Xander Bouwman, Victor Le Pochat, Pawel Foremski, Tom Van Goethem, Carlos H. Gañán, Giovane C. M. Moura, Wouter Joosen, Michel van Eeten. "Helping

hands: Measuring the impact of a large threat intelligence sharing community". 31st USENIX Security Symposium (USENIX Security 2022).

8. Victor Le Pochat, Laura Edelson, Tom Van Goethem, Wouter Joosen, Damon McCoy, Tobias Lauinger. "An Audit of Facebook's Political Ad Policy Enforcement". 31st USENIX Security Symposium (USENIX Security 2022).
Distinguished Paper Award

9. Karel Dhondt, Victor Le Pochat, Alexios Voulimeneas, Wouter Joosen, Stijn Volckaert. "A Run a Day Won't Keep the Hacker Away: Inference Attacks on Endpoint Privacy Zones in Fitness Tracking Social Networks". ACM Conference on Computer and Communications Security (CCS 2022). DOI: 10.1145/3548606.3560616

Publications in peer-reviewed workshop proceedings:

10. Victor Le Pochat, Tom Van Goethem, Wouter Joosen. "A Smörgåsbord of Typos: Exploring International Keyboard Layout Typosquatting". 4th International Workshop on Traffic Measurements for Cybersecurity (WTMC 2019). (Proceedings of the 2019 IEEE Security and Privacy Workshops.) DOI: 10.1109/SPW.2019.00043
Distinguished Paper Award

11. Victor Le Pochat, Tom Van Goethem, Wouter Joosen. "Evaluating the Long-term Effects of Parameters on the Characteristics of the Tranco Top Sites Ranking". 12th USENIX Workshop on Cyber Security Experimentation and Test (CSET 2019).

12. Yana Dimova*, Gertjan Franken*, Victor Le Pochat*, Wouter Joosen, Lieven Desmet. *(* Joint first author.)* "Tracking the evolution of cookie-based tracking on Facebook". 21st Workshop on Privacy in the Electronic Society (WPES 2022). DOI: 10.1145/3559613.3563200

FACULTY OF ENGINEERING SCIENCE
DEPARTMENT OF COMPUTER SCIENCE
DISTRINET
Celestijnenlaan 200A bus 2402
B-3001 Leuven
victor.lepochat@kuleuven.be
https://distrinet.cs.kuleuven.be