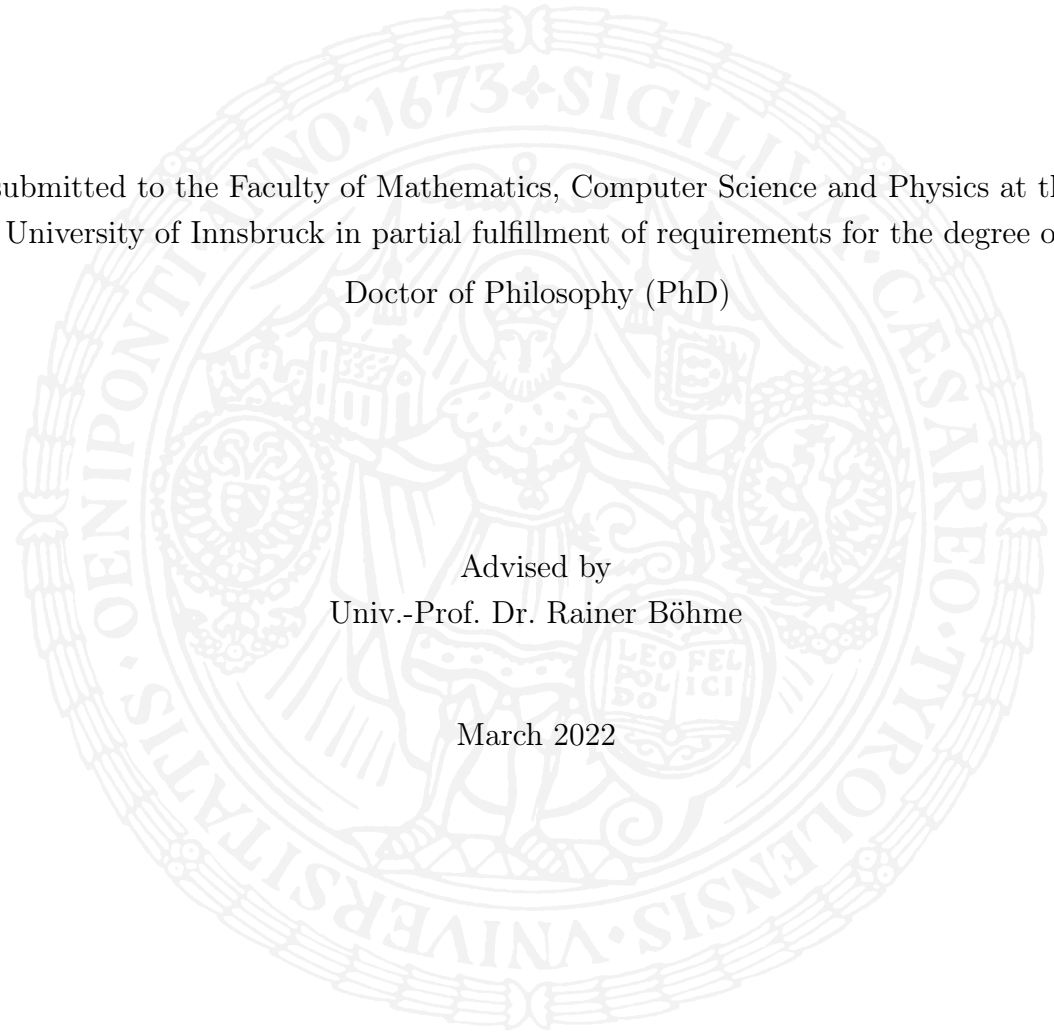
universität innsbruck

Dissertation

Privacy Preference Signals

Maximilian Hils



submitted to the Faculty of Mathematics, Computer Science and Physics at the University of Innsbruck in partial fulfillment of requirements for the degree of

Doctor of Philosophy (PhD)

Advised by
Univ.-Prof. Dr. Rainer Böhme

March 2022

SUMMARY

Privacy preference signals are digital representations of how users want their personal data to be processed. Due to the competing interests of users and data processors, the adoption of such signals remains an unsolved problem despite efforts dating back to the 1990s. The commencement of privacy laws like the EU General Data Protection Regulation (GDPR) in 2018 prompted a new wave of signals as internet firms are pushed to obtain user consent.

This thesis examines the emergence of this consent ecosystem, whose standardized signals now govern many cookie dialogs on the web. Using cross-sectional and longitudinal web measurements, we show which factors drive signal adoption, quantify the impact of signals, and integrate these post-GDPR developments into the wider history of privacy preference signals.

ZUSAMMENFASSUNG

Privacy Preference Signals sind die digitale Repräsentation der Datenschutzpräferenzen eines Nutzers. Auf Grund der gegensätzlichen Interessen von Nutzern und datenverarbeitenden Firmen bleibt die Einführung bzw. Befolgung solcher Signale trotz Bemühungen seit den 1990er-Jahren ein ungelöstes Problem. Das Inkrafttreten von Datenschutzbestimmungen wie der EU Datenschutz-Grundverordnung (GDPR) in 2018 hat dabei eine neue Welle an Signalen ausgelöst, da Firmen nun angehalten sind, die Einwilligung (Consent) von Nutzern zur Datenverarbeitung einzuholen.

Diese Dissertation untersucht die Entstehung des Ökosystems um Consent, dessen standardisierte Signale maßgeblichen Einfluss auf viele Cookie-Dialoge im Internet haben. Basierend auf einmaligen und longitudinalen Web-Messungen zeigt diese Arbeit, welche Faktoren die Einführung von Privacy Preference Signals beeinflussen, welche Auswirkungen das Senden dieser Signale hat, und wie die Entwicklungen nach Inkrafttreten der GDPR im historischen Kontext einzuordnen sind.

ACKNOWLEDGEMENTS

Throughout my PhD, I have received a great deal of support from colleagues, friends, and family members who accompanied me along the way. While it's impossible to list everyone, I would like to highlight a few people.

My journey towards a PhD started in 2013 when I reached out to Rainer Böhme for my Bachelor thesis. A few emails were exchanged, meetings were had, a thesis was written, and I somehow ended up with a position as a student assistant in Münster. Already back then — being the young greenhorn in the group — I received unparalleled support and mentorship from Rainer, which made it easy to follow the group to Innsbruck to pursue PhD in 2017. Over here, I distinctly remember Rainer getting up at 5:30 am on a Saturday to help me polish my first paper for a 7:00 am deadline (I pulled an all-nighter). I hope my time management has gotten a bit better, but I owe you my sincere gratitude for your continuous and unwavering support. I pledge that there are no scaleboxes in this thesis and everything has been Ti*k*Zed.

Staying in the group, I am incredibly thankful for the collaboration with Daniel Woods, with whom I had the pleasure to first share a flat and then form a very successful and symbiotic paper machine. Thank you, Daniel! Of course, my gratefulness extends to all other members of the Security & Privacy Lab. It's been a privilege to be among such a fantastic group of friends, both at work and outside of work.

Outside of Innsbruck, I'd like to thank Aldo Cortesi for his terrific mentorship, which even predates Rainer's involvement (my first mitmproxy commit was in 2012). Netograph's data feeds have been an invaluable foundation for much of the research in this thesis. Thank you, Aldo!

Crossing the Pacific, we arrive at the Good family, my second home in California. Yay and Nathan, thank you for welcoming me into your family during my semester abroad in 2016 and the close friendship thereafter! Building the Berkeley network course in 2018 surely has not accelerated my thesis progress, but it certainly was a part of my PhD journey I do not wish to miss.

Turning back home, all of this wouldn't have been possible without the immense support, guidance, and encouragement from my family. Thank you, Mom and Dad, for raising me to always be curious, ambitious, confident, and humble. You did a fantastic job.

Last but not least, I wouldn't be finished with my PhD yet were it not for a friendly competition on who finishes their thesis first. Thank you, Andrea, for being such a wonderful partner. Winning this contest counts little compared to having you in my life.

CONTENTS

Part I

RESEARCH SUMMARY

1 INTRODUCTION

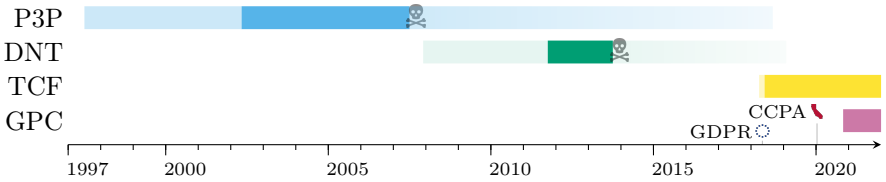


Figure 1: Attempts to standardize privacy preference signals reach back to 1997. A new wave of signals was sparked by EU and US legislation in 2018. Figure adapted from [1, Figure 1].

1.1 THE HISTORY OF PRIVACY PREFERENCE SIGNALS

In 1890, Samuel Warren and Louis Brandeis advocated for the "right to be let alone" in face of increasingly inquisitive journalism: "The press is overstepping in every direction the obvious bounds of propriety and of decency. [...] To satisfy a prurient taste the details of sexual relations are spread broadcast in the columns of the daily papers." [2]. Their essay — *The Right to Privacy*, published in the 1890 Harvard Law Review — is now widely credited with the invention of privacy as a legal concept [3]. 77 years later, Alan Westin's *Privacy and Freedom* expanded the definition of privacy as a legal right for use in modern times [4]. Already in 1965, two years before the first ARPANET computers were connected, Westin warned how the amassing of personal data into gigantic databases threatens individuals [5]. Of course, his definition of privacy as *control over personal information* has become ever more important with the emergence of the internet. Starting with the standardization of cookies in 1997 [6], persistent identifiers that are tied to extensive advertising profiles have become an integral part of the web.

Privacy preference signals are digital representations of how users want their personal data to be processed. These signals vary from binary "Do Not Track" signals through to more complex expressions in cookie consent dialogs. While users may send particular signals to limit how their own personal data is processed, companies may also proactively collect privacy preferences to legitimize their own data processing (as it may be required by law). In 2022, the most common manifestation of this are modern cookie banners and consent dialogs (see Appendix A for examples). However, the efforts to provide internet users with means to express their privacy preferences go back to at least the Platform for Privacy Preferences (P3P) (see Figure 1). P3P was standardized by the World Wide Web Consortium (W3C) in 2002, was adopted by around 20k websites [7], but eventually succumbed to a slow death starting in 2007. Reasons for its decline included the lack of consequences for false reporting of

	P3P	DNT	TCF	GPC
Proposed	1997	2009	2018	2020
Legal Basis	–	–	GDPR	CCPA
Design				
Signal	complex	1 bit	complex	1 bit
Impl.	Policy XML	HTTP Header	Consent Dialog	HTTP Header
Adoption				
Websites	> 20k [11, 12, 7, 13]	> 9 [14]	8 265*	?
AdTech	> 11 [7]	≈ 0 [14]	684*	?
Browsers				

? = unknown, ✅ = compat. with exist. tech., * = own measurements, see Sec. 4.5

Table 1: Different privacy signals vary in their design: While P3P and TCF propose complex frameworks with fine-grained privacy settings, DNT and GPC simplify user choice to a single bit. Table adapted from [1, Table 1].

privacy practices as well as the complexity of the specification [8]. Another W3C working group was then formed to specify the Do Not Track (DNT) standard [9], which was conceptually much simpler: Users communicated their desire that no data regarding their activity should be collected or shared using a single-bit HTTP header (see Figure 1). Again, the standard failed as the working group was closed before completion, citing a lack of planned support across "the ecosystem at large" [10].

DNT's proponents found out the hard way that adopting a privacy preference signal is a coordination problem: While browsers (*user* agents) adopted P3P and DNT, vendors profiting from personal data were not incentivized to adopt these standards and respect the wishes expressed by data subjects [15]. If we want to understand why current privacy preference signals may succeed or fail, we thus need to look at why previous approaches failed. In the first paper in this thesis (Chapter 4), I provide a history of the privacy preference signals introduced here and analyze how the more recent TCF signal – subject of the next subsections – achieved more widespread adoption.

Skeptics will argue that privacy preference signals only provide *soft privacy*, i.e., they do not put any technical measures in place that could stop the receiving party from disregarding users' preferences and processing personal data nonetheless. Users must trust the signal's recipient to be compliant and stop processing personal data. In contrast, *hard privacy* technologies such as encryption would provide technical guarantees that user data is protected [16]. When looking at past behavior of AdTech companies, we will however find that soft privacy may not be the toothless tiger some critics make it out to be [17]. With both P3P and DNT, AdTech companies proactively invested resources in the W3C working groups to influence the respective signal design. The DNT mailing list archives hold more than 11,000 messages [18]. As the standard started to pose a threat to AdTech — Microsoft announced it would be turned on by default —, the Interactive Advertising Bureau (IAB), a coalition of advertisers, immediately withdrew from the working group to delegitimize the

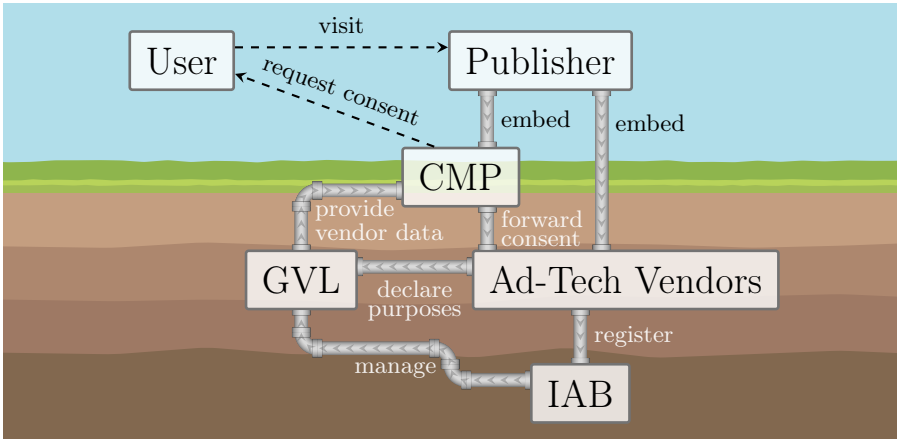


Figure 2: Surfacing the web's new compliance engine: Publishers embed Consent Management Providers (CMPs), which display consent prompts to users and forward consent decisions to ad-tech vendors. In the background, the Interactive Advertising Bureau (IAB) orchestrates this through its Transparency and Consent Framework (TCF). Figure adapted from [24, Figure 2].

standardization process [15]. TCF was developed via a working group that is managed by the IAB and lists 156 participating organizations [19]. On the opposing side, privacy advocates have taken the trouble to file more than 500 complaints against misleading cookie banners in 2021 [20]. On a policy level, the design of privacy preference signals has been subject to committee hearings at the US Congress [21, 22] and the EU parliament [23]. All this speaks to the practical importance of soft privacy and privacy preference signals.

1.2 GDPR AND THE RISE OF CONSENT MANAGEMENT

The passage of the EU General Data Protection Regulation (GDPR) and the California Consumer Privacy Act (CCPA) set off a second wave of privacy preference signals starting in 2018. In the EU, the GDPR now establishes that websites need their users' "freely given, specific, informed and unambiguous" consent for certain data processing purposes [25]. In California, CCPA now requires websites to allow users to opt-out of the sale of their personal data [26]. Virginia's Consumer Data Protection Act, set to go into effect in 2023, will require businesses to obtain opt-in consent to process sensitive data [27]. Leaving the Western Hemisphere, Brazil's Data Protection Bill of Law already establishes regulations similar to the GDPR in Brazil [28]. So does Jamaica's Data Protection Act of 2020 [29]. To comply with these laws, technical infrastructure to acquire consent must be designed so that websites and AdTech vendors can continue to process personal data.

In the past, each website offered a unique privacy policy and cookie notice [30, 31]. This diversity overwhelmed users who could not commit hundreds of hours to read each policy [32, 33] nor navigate novel interface designs without making errors [34]. While P3P and DNT would have provided common browser interfaces to remedy the problem, AdTech resisted adopting a standard

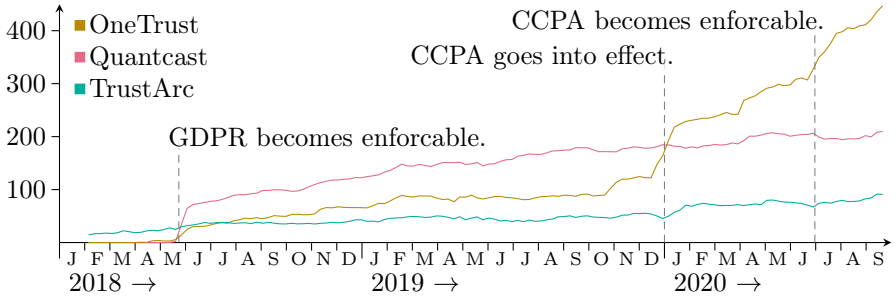


Figure 3: Number of websites in the Tranco 10k toplist that have adopted popular consent management providers. We can observe spikes in adoption around the enforcement dates for GDPR (May 2018) and CCPA (Jan/Jul 2020). Figure adapted from [24, Figure 6].

that would have made their data collection more difficult. The GDPR's new imperative to manage and document consent turned the existing heterogeneity into a potential liability as AdTech vendors would need to rely on individual websites to properly collect consent. In reaction to the GDPR, the IAB Europe set out to develop their own privacy preference signal, the Trust and Consent Framework (TCF). In contrast to previous standards, membership in the working group is being controlled by IAB Europe, and the standard is predominantly developed by private firms from the advertising and publishing industries [35]. The IAB fittingly describes TCF as "the only GDPR consent solution built by the industry for the industry" [36].

The TCF defines the legal terms and data processing purposes that users consent to and the format by which consent signals are stored and exchanged between websites and AdTech companies. It is implemented by websites in the form of a consent dialog that is technically part of the webpage. It does not require browser vendors to adopt the signal, as was the case with DNT. Instead, it creates the role of *Consent Management Providers (CMPs)*, who implement the framework on individual websites. CMPs are central to the TCF in providing an interface between website, user, and ad vendors. They provide websites with a (customizable) cookie prompt to embed, store users' choices as browser cookies, and provide an API for advertisers to access this information (see Figure 2).

To receive TCF consent signals from CMPs, AdTech vendors must register with the IAB and pay a yearly maintenance fee of 1,500€ to join the *Global Vendor List (GVL)*. As of January 2022, 784 companies are registered on this list. Most CMPs collect consent for the entire GVL by default, which allows them to share consent decisions across multiple websites [37].

While P3P and DNT required active adoption from a party that was generally disincentivized to do so (AdTech), TCF could be adopted by websites with little recourse from browsers or users. The increasing complexity of the legal landscape and uncertainty around sanctions for non-compliance led many websites to adopt CMPs (see Figure 3). After custom cookie banners on websites [38] and abandoned standards like P3P and DNT, the successful rise of TCF represents a new stage in how privacy preferences are communicated.

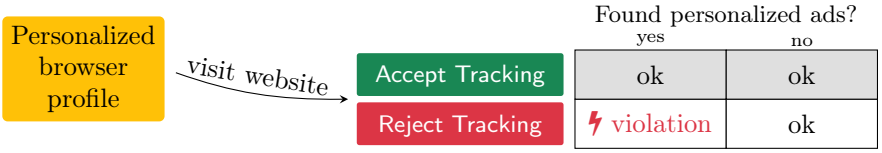


Figure 4: To measure the impact of consent dialogs on ad personalization, a series of websites is visited with a primed browser profile. On each website, the user accepts or rejects all tracking. If no consent was given, the presence of personalized ads is taken as an indicator for unlawful data processing. Figure adapted from [39, Figure 1].

In the second paper of this thesis (Chapter 5), I measure the formation of this ecosystem using longitudinal measurements. Based on 161 million browser crawls, the paper provides evidence that CMP adoption is driven by the enforcement of privacy laws, and that a few CMPs have set out to dominate the market. Although the market power of CMPs and the dominance of the IAB in the design of the privacy preference signal is worrying, the same standardization opens up novel measurement opportunities.

1.3 THE EFFECT OF PRIVACY PREFERENCE SIGNALS

A key question with privacy preference signals is whether third parties are compliant and stop processing personal data when instructed to do so (see *soft privacy* in Section 1.1). Researchers have tried to answer this question by checking whether the user's browser transmits the correct TCF signal in its HTTP requests to third parties [40, 41]. While this method uncovers obvious privacy violations where the website owner already misrepresents the user's decision, it does not help to understand whether the embedded AdTech vendors are compliant when receiving a correct TCF signal. AdTech vendors have monetary incentives to build extensive ad profiles, which is at odds with respecting the user's wishes. To provide some anecdotal evidence: Pesch interviewed advertising companies and found that some only joined the IAB's GVL because business partners required membership; they claimed their own data processing would not require consent [42]. Clearly, if compliance of vendors cannot be measured, they have little incentive to actually stop amassing the gigantic databases Westin warned about (see Section 1.1).

Measuring whether AdTech partners respect TCF signals is more tricky as researchers cannot scrutinize AdTech's server-side processing code. One possible approach to solve this problem is to perform end-to-end measurements (see Figure 4). Users first prime a browser profile with specific interests by visiting related websites and searching for relevant terms. This places cookies and other tracking identifiers in their browser profile which can then be picked up by AdTech. In a second step, they visit generic websites where they either accept or reject all tracking in the consent dialog (the treatment factor in our study design). If no consent is given to personalize ads and all vendors behave correctly, advertisements for the primed interests should be highly unlikely to

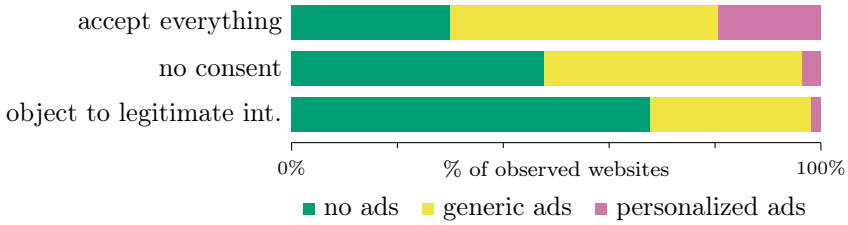


Figure 5: Observed ads depending on the communicated privacy preference signal. Manual measurements show that not providing consent in TCF dialogs removes most ads personalization on a sample of news websites. When additionally objecting to data processing based on legitimate interests, many publishers opt to not show any ads at all ($N = 44$, Nov. 2022). Figure adapted from [39, Figure 5].

appear. Seeing a significant amount of personalized ads would be an indication of misconduct by the embedded AdTech parties.

In the third paper of this thesis, I performed such end-to-end measurements both manually and in an automated fashion. Using 44 manual measurements performed by students, the paper shows that rejecting a consent dialog does indeed stop the majority of ad personalization (see Figure 5). Curiously, when users are instructed to also object to data processing based on legitimate interests (an advanced — usually well-hidden — opt-out mechanism available in TCF consent dialogs), a large fraction of websites even opt to not show any advertisements. This behavior is difficult to explain as the same websites were capable of showing non-personalized ads when the user simply clicked "Reject All" without explicitly objecting to legitimate interests.

Scaling this manual study to a large sample set by automating the priming process, the consent dialog interaction, and the personalization measurement turned out to be more difficult than expected. Using the automated approach it was only possible to measure comparably very small effect sizes, which can be attributed to AdTech's strong anti-bot measures. This leaves this part of the paper with a meta result: Measuring the effects of privacy preference signals at scale is a very hard problem to solve. This is bad news for privacy advocates and data protection agencies, who need to rely on laborious manual methods to keep tabs on AdTech. However, despite these limitations, the manual analysis in this paper has shown that consent decisions are widely respected by major players in the ecosystem.

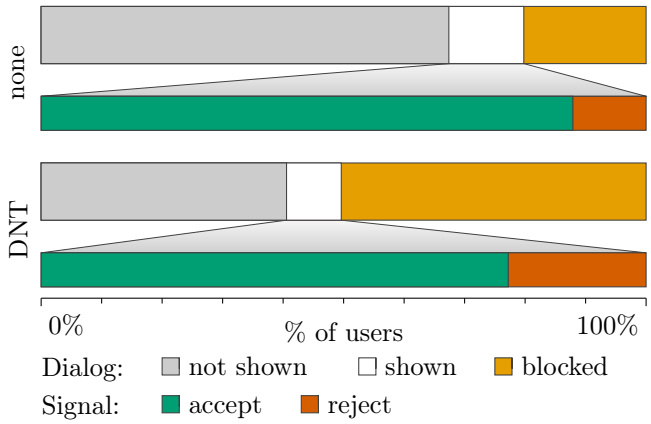


Figure 6: Users that send a Do Not Track (DNT) header are more likely to block cookie dialogs or refuse consent when presented with one. However, more than 75% of DNT users still click "Accept". This creates ambiguous privacy preference signals as users technically provide consent via TCF but also indicate that they do not like to be tracked via DNT. Figure adapted from [43, Figure 3].

1.4 CONFLICTING SIGNALS

In theory, privacy laws like GDPR and CCPA should empower users to control how their personal data is processed. However, the adoption of TCF and the failure of competing signals have made it difficult and time-consuming for users to exercise their rights. As described in the second paper of this thesis, AdTech unsurprisingly put their own interests first in the design of the TCF standard.

With TCF succeeding, the competing interests of privacy advocates led to the proposal of a new standard, the Global Privacy Control (GPC) [44]. GPC, first released as an unofficial draft specification in October 2020, aims to provide a simpler way to exercise one's rights. Last updated in January 2022, it follows the spirit of DNT in using a single-bit HTTP header to communicate preferences. However, while DNT lacked means of legal enforcement, GPC reframes "Do Not Track" as a "Do Not Sell" request under CCPA, and as a general request not to sell or share data with other data controllers under GDPR [45]. This means that in contrast to DNT, GPC references specific enforceable legal rights. This, GPC proponents argue, makes it possible to enforce that AdTech companies adopt and respect the signal.

An important difference between DNT/GPC and TCF is that they do not operate on the same technical layer. While TCF is directly embedded into websites, DNT and GPC are sent by the browser or by browser extensions. This implies that users may transmit more than one signal at the same time and thereby express conflicting or ambiguous preferences. For example, a user may have configured their browser to always send a GPC "Do Not Sell" signal, but at the same time they may click "Accept" on a TCF cookie dialog embedded on a website.

The possibility of users sending multiple signals raises questions about legal interpretation under GDPR, CCPA, or other privacy laws. For example, one

could argue that TCF signals should take precedence because their consent is given specifically for an individual website the user trusts, whereas DNT/GPC is a global non-specific browser setting. On the other hand, existing research has shown that TCF dialogs often employ deceptive design patterns [46], which puts the validity of TCF "consent" into question. More concretely, the GDPR specifically stipulates that consent must be *unambiguous*. It is clear that the question of which signal takes precedence under such circumstances can only be answered by legal analysis. However, such analysis first needs to be supported by showing that the problem is not theoretical and users do send conflicting privacy preference signals in the wild. In the final paper of this thesis (Chapter 7), I show that conflicting signals do exist in the wild. The paper finds a sizable number of users who manually enabled DNT in their browser settings, but who nonetheless accept a TCF consent dialog (see Figure 6). Additionally, the data also shows that TCF consent dialogs are often blocked entirely by adblocking extensions. This phenomenon was previously overlooked in the academic discourse and showcases another approach to how users may indirectly express their privacy preferences in a way that is not prescribed by AdTech.

Section 2 provides the reader with a short overview of all papers in this thesis. In Section 3, I discuss the most important open questions with regard to privacy preference signals and conclude the dissertation. Part II contains the four papers included in this thesis.

2

SUMMARY OF PAPERS

Over the course of my PhD, I have (co-)authored five papers, four of which are included in this dissertation. To provide the reader with a high-level overview, I shortly summarize each paper and my personal contribution to it in this section. Where available, the acceptance rate and CORE rank [47] of the conference is provided as additional context. My contribution to each paper is evaluated in three categories: i) the conception of the research idea, ii) the operational research work, and iii) the write-up of the paper.

Maximilian Hils, Daniel W. Woods, and Rainer Böhme. Privacy Preference Signals: Past, Present and Future. Proceedings on Privacy Enhancing Technologies*, (4), 2021.* https://doi.org/10.2478/popets-2021-0069

CORE Rank: A
Acceptance Rate: 19%

ABSTRACT. Privacy preference signals are digital representations of how users want their personal data to be processed. Such signals must be adopted by both the sender (users) and intended recipients (data processors). Adoption represents a coordination problem that remains unsolved despite efforts dating back to the 1990s. Browsers implemented standards like the Platform for Privacy Preferences (P3P) and Do Not Track (DNT), but vendors profiting from personal data faced few incentives to receive and respect the expressed wishes of data subjects. In the wake of recent privacy laws, a coalition of AdTech firms published the Transparency and Consent Framework (TCF), which defines an opt-in consent signal. This paper integrates post-GDPR developments into the wider history of privacy preference signals. Our main contribution is a high-frequency longitudinal study describing how TCF signal gained dominance as of February 2021. We explore which factors correlate with adoption at the website level. Both the number of third parties on a website and the presence of Google Ads are associated with higher adoption of TCF. Further, we show that vendors acted as early adopters of TCF 2.0 and provide two case studies describing how Consent Management Providers shifted existing customers to TCF 2.0. We sketch ways forward for a pro-privacy signal.

CONTRIBUTION (40%, 70%, 20%). This paper stands out from the rest in that its combination of qualitative aspects (describing past signals) and quantitative aspects (longitudinal measurements of TCF) required several rounds of revisions and additional measurements. The feedback we received from the anonymous PETS reviewers was exceptionally constructive and led to a paper that now comprehensively covers the history of privacy preference signals. My main contribution to this study is the measurement of TCF 2's uptake using multiple methods, which we collaboratively integrated into the wider history of privacy preference signals.

Maximilian Hils, Daniel W Woods, and Rainer Böhme. Measuring the Emergence of Consent Management on the Web. In Proceedings of the Internet Measurement Conference 2020*, IMC '20. ACM, 2020.* https://doi.org/10.1145/3419394.3423647

CORE Rank: A
Acceptance Rate: 24.5%

ABSTRACT. Privacy laws like the General Data Protection Regulation (GDPR) and the California Consumer Privacy Act (CCPA) have pushed internet firms processing personal data to obtain user consent. Uncertainty around sanctions for non-compliance led many websites to embed a Consent Management Provider (CMP), which collects users' consent and shares it with third-party vendors and other websites. Our paper maps the formation of this ecosystem using longitudinal measurements. Primary and secondary data sources are used to measure each actor within the ecosystem. Using 161 million browser crawls, we estimate that CMP adoption doubled from June 2018 to June 2019 and then doubled again until June 2020. Sampling 4.2 million unique domains, we observe that CMP adoption is most prevalent among moderately popular websites (Tranco top 50-10k) but a long tail exists. Using APIs from the ad-tech industry, we quantify the purposes and lawful bases used to justify processing personal data. A controlled experiment on a public website provides novel insights into how the time-to-complete of two leading CMPs' consent dialogues varies with the preferences expressed, showing how privacy aware users incur a significant time cost.

CONTRIBUTION (60%, 95%, 60%). This paper is included as the second publication in my thesis, yet it was written before the first one. While I laid the foundation for many of the PETS paper's measurements here, it only makes sense to start the narrative with the history of privacy preference signals. After collaboratively developing the research idea, my individual contribution in this work is the conception, implementation, validation, and technical description of our longitudinal measurement instrument.

Maximilian Hils. Measuring the Impact of Privacy Preference Signals, 2022. (work-in-progress, to be submitted)

ABSTRACT. Since the passage of the General Data Protection Regulation (GDPR) in Europe, many websites employ cookie dialogs to obtain consent from users. Previous research has shown that AdTech regularly uses dark patterns in these dialogs to trick users into consenting. This paper goes beyond the user interface and sets out to analyze whether clicking "Reject All" in a cookie dialog does actually stop the data processing.

We perform manual and automated end-to-end measurements in which we first create personalized browser profiles, and then measure how different consent signals affect observed ad personalization. Our user study with 2093 website observations shows that many AdTech providers do indeed respect negative signals and stop showing personalized ads. We attempt to automate our measurements and instrument major browser engines from different vantage points using multiple crawling strategies. However, we find that the effects of privacy preference signals are hard to measure at scale due to AdTech's anti-bot measures.

While our main result is a positive one (AdTech respecting privacy preference signals), we suggest that this is simply because the risk of non-compliance currently outweighs the profit that could be gained from the small minority of users who do not give consent. With regulators enforcing easier opt-out mechanisms, measuring compliance will become increasingly necessary.

CONTRIBUTION (100%, 100%, 100%). Building on the experience gained from the other papers in my thesis, I have carried out all work for this study on my own. Compared with the other papers it has the most ambitious measurement setup, examining compliance under the TCF from a novel perspective.

Maximilian Hils, Daniel W. Woods, and Rainer Böhme. Conflicting Privacy Preference Signals in the Wild. In Data Protection and Privacy*, volume 15, 2022. To appear.*

ABSTRACT. Privacy preference signals allow users to express preferences over how their personal data is processed. These signals become important in determining privacy outcomes when they reference an enforceable legal basis, as is the case with recent signals such as the Global Privacy Control and the Transparency & Consent Framework. However, the coexistence of multiple privacy preference signals creates ambiguity as users may transmit more than one signal. This paper collects evidence about ambiguity flowing from the aforementioned two signals and the historic Do Not Track signal. We provide the first empirical evidence that ambiguous signals are sent by web users in the wild. We also show that preferences stored in the browser are reliable predictors of privacy preferences expressed in web dialogs. Finally, we provide the first evidence that popular cookie dialogs are blocked by the majority of users who adopted the Do Not Track and Global Privacy Control standards. These empirical results inform forthcoming legal debates about how to interpret privacy preference signals.

CONTRIBUTION (80%, 100%, 50%). This paper completes my thesis with a user study that I designed and performed to investigate the occurrence of conflicting privacy preference signals. It builds on a body of existing GDPR user studies, but introduces additional factors such as the influence of adblockers (which were not considered previously) into the academic discourse. I contributed significantly to the conception of the research question, the write-up, and all other parts of the research process. In particular, the writing of the paper made me pick up a succinct style to concisely communicate method and results given a short page limit.

3

CONCLUSION

This dissertation has covered privacy preference signals on the web from a variety of perspectives. For these signals to be successful, they must be adopted by both senders (users) and recipients (AdTech). However, as outlined in the first paper of this thesis, these two parties tend to have very different interests. On the one side, AdTech vendors and publishers don't want to forego their highly lucrative targeted advertising practices. This first led them to send intentionally misconfigured P3P policies, and then later quit the "Do Not Track" standardization process once it threatened to be on by default. On the other side, users would like simple and effective privacy preference signals to stop rampant and intransparent data collection by third parties. It's evident that consensus-based standardization is doomed to fail in the face of these conflicting interests. The unfortunate conclusion for technologists is that the success of future privacy preference signals will likely not be determined by their technical merit.

Given the hardened fronts between privacy-aware users and AdTech, we naturally arrive at the question of whether regulators can resolve the conflict. One well-intentioned attempt at this was the passage of the GDPR, forcing AdTech to either obtain consent or dial back on data collection. However, faced with this choice, AdTech developed the TCF signal — serving their own interests only —, and unilaterally forced it upon users in the form of consent dialogs on websites. This widespread emergence of "cookie banner terror" [20], tracked in the second paper of this thesis, is certainly not what regulators had in mind. Resolving the situation will require future interventions to reduce the number of decisions users need to make.

Putting regulation aside, a key problem with privacy preference signals remains the lack of hard privacy. Users need to trust AdTech to honor preference signals over its own financial interests. It is unrealistic to expect AdTech to concede voluntarily if no one is watching, so some enforcement needs to happen. The third paper in this thesis shows that measuring compliance — a prerequisite for enforcement — is already difficult from a technical perspective. Consolation may come from the idea that larger players in the ecosystem are under closer public scrutiny, leaving fewer opportunities to disregard signals. Nonetheless, the IAB's vendor list has 794 entries. It is unclear how compliance can be enforced in the dark alleys of AdTech's tracking ecosystem.

Looking ahead, can we fix the consent dialog mess that the GDPR inadvertently created? The rising popularity of new privacy signals such as GPC may be reason for cautious optimism. However, for user-friendly signals to succeed internationally, regulators must establish strong practical precedent on conflicting signals. AdTech will argue that their site-specific consent dialogs should take priority over global permanent signals like GPC. Unless regulators take a strong stance against this interpretation, consent dialogs and dark patterns won't go away.

REFERENCES

[1] Maximilian Hils, Daniel W. Woods, and Rainer Böhme. Privacy Preference Signals: Past, Present and Future. *Proceedings on Privacy Enhancing Technologies*, (4), 2021. https://doi.org/10.2478/popets-2021-0069.

[2] Samuel D. Warren and Louis D. Brandeis. The Right to Privacy. *Harvard Law Review*, 4(5):193–220, 1890. https://www.jstor.org/stable/1321160.

[3] Dorothy J. Glancy. The Invention of The Right to Privacy. *Arizona Law Review*, 21(1):1–39, 1979. https://digitalcommons.law.scu.edu/facpubs/317/.

[4] Alan F. Westin. *Privacy and Freedom*. Atheneum, New York, 1967.

[5] Daniel J. Solove. *Introduction to Alan F. Westin's Privacy and Freedom*. Ig Publishing, 2015. ISBN 9781632460738.

[6] David Kristol and Lou Montulli. *RFC 2109 – HTTP State Management Mechanism*. Internet Engineering Task Force, 1997. https://www.ietf.org/rfc/rfc2109.

[7] Pedro Giovanni Leon, Lorrie Faith Cranor, Aleecia M McDonald, and Robert McGuire. Token attempt: The misrepresentation of website privacy policies through the misuse of P3P compact policy tokens. In *ACM Workshop on Privacy in the Electronic Society*, pages 93–104, 2010. https://doi.org/10.1145/1866919.1866932.

[8] Electronic Privacy Information Center and Junkbusters. Pretty Poor Privacy: An Assessment of P3P and Internet Privacy. https://epic.org/reports/prettypoorprivacy.html, 2000.

[9] World Wide Web Consortium. Tracking Protection Working Group. https://www.w3.org/2011/tracking-protection/, 2011.

[10] Tracking Protection Working Group. WG closed. https://github.com/w3c/dnt/commit/5d85d6c, 2019.

[11] Simon Byers, Lorrie Faith Cranor, and David Kormann. Automated analysis of P3P-enabled web sites. In *Proceedings of the 5th International Conference on Electronic Commerce*, pages 326–338, 2003. https://doi.org/10.1145/948005.948048.

[12] Patricia Beatty, Ian Reay, Scott Dick, and James Miller. P3P adoption on e-Commerce web sites: a survey and analysis. *IEEE Internet Computing*, 11(2):65–71, 2007. https://doi.org/10.1109/MIC.2007.45.

[13] Ian Reay, Patricia Beatty, Scott Dick, and James Miller. Privacy policies and national culture on the internet. *Information Systems Frontiers*, 15 (2):279–292, 2013. https://doi.org/10.1007/s10796-011-9336-7.

[14] Future of Privacy Forum. Companies that have implemented Do Not Track. https://allaboutdnt.com/companies/, 2020.

[15] Interactive Advertising Bureau. "Do Not Track" set to "On" by Default in Internet Explorer 10—IAB Response. https://www.iab.com/news/do-not-track-set-to-on-by-default-in-internet-explorer-10iab-response/, 2012.

[16] Mina Deng, Kim Wuyts, Riccardo Scandariato, Bart Preneel, and Wouter Joosen. A privacy threat analysis framework: supporting the elicitation and fulfillment of privacy requirements. *Requirements Engineering*, 16(1): 3–32, November 2010. https://doi.org/10.1007/s00766-010-0115-7.

[17] George Danezis. Introduction to privacy technology. Talk at COSIC, Katholieke University Leuven, 2007. http://www0.cs.ucl.ac.uk/staff/G.Danezis/talks/Privacy_Technology_cosic.pdf.

[18] World Wide Web Consortium (W3C). public-tracking@w3.org Mail Archives, 2022. https://lists.w3.org/Archives/Public/public-tracking/.

[19] IAB Tech Lab. Global Privacy Working Group. https://iabtechlab.com/working-groups/global-privacy-working-group/, 2022.

[20] NOYB – European Center for Digital Rights. noyb aims to end "cookie banner terror" and issues more than 500 GDPR complaints, 2021. https://noyb.eu/en/noyb-aims-end-cookie-banner-terror-and-issues-more-500-gdpr-complaints.

[21] US Senate Committee on Commerce, Science, and Transportation. Hearing on Protecting Consumer Privacy, 2021. https://www.commerce.senate.gov/2021/9/protecting-consumer-privacy.

[22] Committee on the Judiciary. Hearing on GDPR & CCPA: Opt-ins, Consumer Control, and the Impact on Competition and Innovation, 2019. https://www.judiciary.senate.gov/meetings/gdpr-and-ccpa-opt-ins-consumer-control-and-the-impact-on-competition-and-innovation.

[23] European Parliament Committee on Civil Liberties, Justice and Home Affairs. General Data Protection Regulation implementation, enforcement and lessons learned, 2022. https://www.europarl.europa.eu/committees/en/product/product-details/20220301CHE09983.

[24] Maximilian Hils, Daniel W Woods, and Rainer Böhme. Measuring the Emergence of Consent Management on the Web. In *Proceedings of the*

Internet Measurement Conference 2020, IMC '20. ACM, 2020. https://doi.org/10.1145/3419394.3423647.

[25] European Union. Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation). *Official Journal L110*, 59:1–88, 2016-05-04. https://eur-lex.europa.eu/eli/reg/2016/679/.

[26] California Civil Code § 1798. California Consumer Privacy Act of 2018. https://leginfo.legislature.ca.gov/faces/billTextClient.xhtml?bill_id=201720180AB375.

[27] Code of Virginia Title 59.1 Chapter 53. Virginia Consumer Data Protection Act of 2021. https://law.lis.virginia.gov/vacode/title59.1/chapter53/.

[28] Brazilian General Personal Data Protection Law (Lei Geral de Proteção de Dados Pessoais, LGPD). Law No. 13.709/2018. https://www.planalto.gov.br/ccivil_03/_ato2015-2018/2018/lei/l13709.htm.

[29] The Data Protection Act, 2020 (Jamaica). https://japarliament.gov.jm/attachments/article/339/The%20Data%20Protection%20Act,%202020.pdf.

[30] Ryan Amos, Gunes Acar, Eli Lucherini, Mihir Kshirsagar, Arvind Narayanan, and Jonathan Mayer. Privacy Policies over Time: Curation and Analysis of a Million-Document Dataset. In *Proceedings of the Web Conference 2021*. ACM, apr 2021. https://doi.org/10.1145/3442381.3450048.

[31] Martin Degeling, Christine Utz, Christopher Lentzsch, Henry Hosseini, Florian Schaub, and Thorsten Holz. We Value Your Privacy ... Now Take Some Cookies: Measuring the GDPR's Impact on Web Privacy. In *Proceedings 2019 Network and Distributed System Security Symposium*. Internet Society, 2019. https://doi.org/10.14722/ndss.2019.23378.

[32] Aleecia M McDonald and Lorrie Faith Cranor. The cost of reading privacy policies. *Journal of Law and Policy for the Information Society*, 4:543, 2008. ISSN 2372-2959.

[33] Joseph Bonneau and Sören Preibusch. The Privacy Jungle: On the Market for Data Protection in Social Networks. In *8th Annual Workshop on the Economics of Information Security, WEIS*, 2009. https://doi.org/10.1007/978-1-4419-6967-5_8.

[34] Idris Adjerid, Alessandro Acquisti, Laura Brandimarte, and George Loewenstein. Sleights of Privacy: Framing, Disclosures, and the Limits

of Transparency. In *Proceedings of the Ninth Symposium on Usable Privacy and Security*, SOUPS 13. ACM, 2013. https://doi.org/10.1145/2501604.2501613.

[35] IAB Tech Lab. Global Privacy Working Group. https://iabtechlab.com/working-groups/global-privacy-working-group/, 2011.

[36] IAB Europe. What is the Transparency and Consent Framework (TCF)? https://iabeurope.eu/transparency-consent-framework/, 2020.

[37] Daniel W Woods and Rainer Böhme. The Commodification of Consent. In *20th Annual Workshop on the Economics of Information Security, WEIS*, 2020. https://doi.org/10.1016/j.cose.2022.102605.

[38] Martino Trevisan, Stefano Traverso, Eleonora Bassi, and Marco Mellia. 4 Years of EU Cookie Law: Results and Lessons Learned. *Proceedings on Privacy Enhancing Technologies*, 2019(2):126 – 145, 2019. https://doi.org/10.2478/popets-2019-0023.

[39] Maximilian Hils. Measuring the Impact of Privacy Preference Signals, 2022. (work-in-progress, to be submitted).

[40] Célestin Matte, Nataliia Bielova, and Cristiana Santos. Do Cookie Banners Respect my Choice? Measuring Legal Compliance of Banners from IAB Europe's Transparency and Consent Framework. In *IEEE Symposium on Security and Privacy*, pages 791–809. IEEE, 2020. https://doi.org/10.1109/SP40000.2020.00076.

[41] Koen Aerts. Cookie Dialogs and Their Compliance. Master's thesis, Open University of the Netherlands, July 2021. https://www.open.ou.nl/hjo/supervision/2021-koen-aerts-msc-thesis.pdf.

[42] Paulina Jo Pesch. Drivers and Obstacles for the Adoption of Consent Management Solutions by Ad-Tech Providers. In *Euro S&P Workshop on Consent Management in Online Services (COnSeNT)*, 2021. https://doi.org/10.1109/EuroSPW54576.2021.00034.

[43] Maximilian Hils, Daniel W. Woods, and Rainer Böhme. Conflicting Privacy Preference Signals in the Wild. In *Data Protection and Privacy*, volume 15, 2022. To appear.

[44] Ashkan Soltani and Sebastian Zimmeck. Global Privacy Control (GPC). https://globalprivacycontrol.org/.

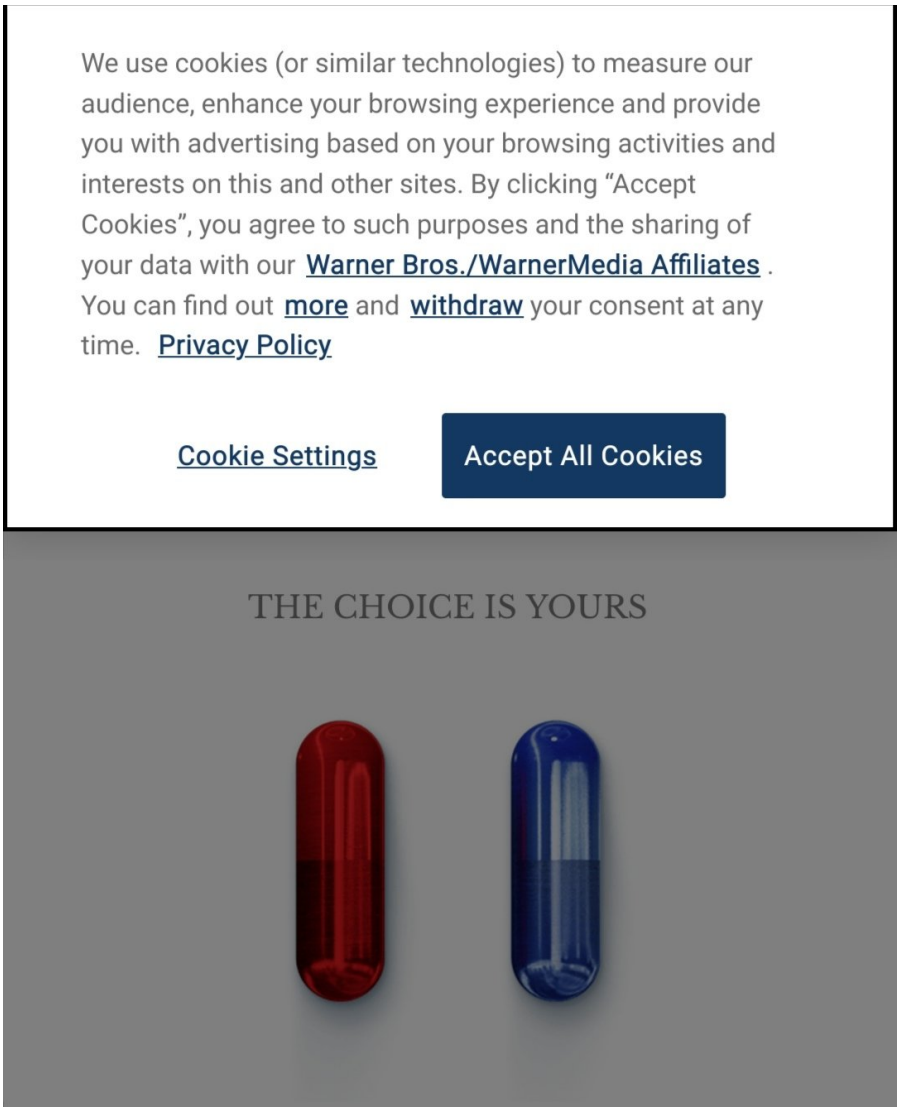
[45] Robin Berjon, Sebastian Zimmeck, Ashkan Soltani, David Harbage, and Peter Synder. Global Privacy Control (GPC) Unofficial Draft 27 January 2022. https://globalprivacycontrol.github.io/gpc-spec/.

[46] Dominique Machuletz and Rainer Böhme. Multiple Purposes, Multiple Problems: A User Study of Consent Dialogs after GDPR. *Proceedings on Privacy Enhancing Technologies*, (2):481–498, 2020. https://doi.org/10.2478/popets-2020-0037.

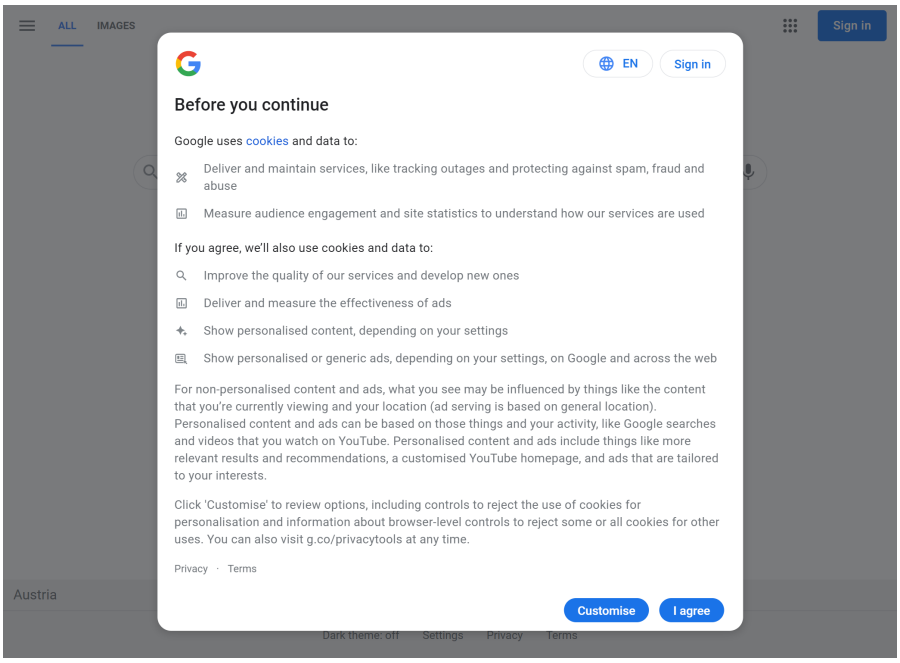
[47] The Computing Research and Education Association of Australasia. CORE Rankings Portal. http://portal.core.edu.au/conf-ranks/, 2021.

APPENDIX

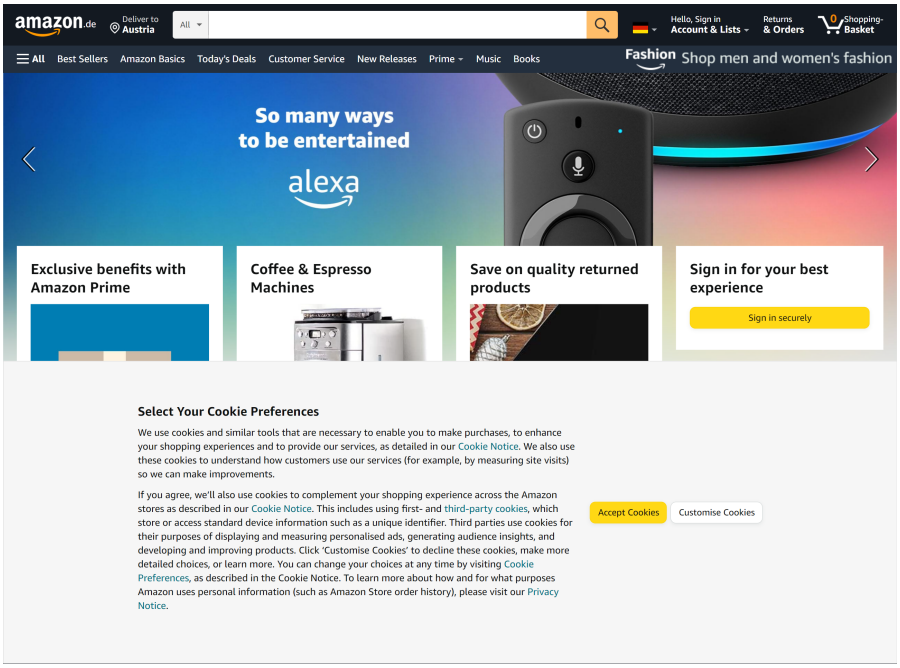
A) CONTEMPORARY CONSENT BANNERS AND COOKIE DIALOGS



Cookie dialog for `thechoiceisyours.whatisthematrix.com` (Sept. 2021)



Cookie dialog for `www.google.com` (Feb. 2022)



Consent banner for `www.amazon.de` (Feb. 2022)

Part II

PAPERS

4

PRIVACY PREFERENCE SIGNALS: PAST, PRESENT AND FUTURE

AUTHORS

Maximilian Hils, University of Innsbruck
Daniel Woods, University of Innsbruck
Rainer Böhme, University of Innsbruck

CONFERENCE

The 21st Privacy Enhancing Technologies Symposium (PETS)
12–16 July 2021, Virtual Event.

ABSTRACT

Privacy preference signals are digital representations of how users want their personal data to be processed. Such signals must be adopted by both the sender (users) and intended recipients (data processors). Adoption represents a coordination problem that remains unsolved despite efforts dating back to the 1990s. Browsers implemented standards like the Platform for Privacy Preferences (P3P) and Do Not Track (DNT), but vendors profiting from personal data faced few incentives to receive and respect the expressed wishes of data subjects. In the wake of recent privacy laws, a coalition of AdTech firms published the Transparency and Consent Framework (TCF), which defines an opt-in consent signal. This paper integrates post-GDPR developments into the wider history of privacy preference signals. Our main contribution is a high-frequency longitudinal study describing how TCF signal gained dominance as of February 2021. We explore which factors correlate with adoption at the website level. Both the number of third parties on a website and the presence of Google Ads are associated with higher adoption of TCF. Further, we show that vendors acted as early adopters of TCF 2.0 and provide two case-studies describing how Consent Management Providers shifted existing customers to TCF 2.0. We sketch ways forward for a pro-privacy signal.

4.1 INTRODUCTION

Privacy preference signals are digital representations of how users want their personal data to be processed. These vary from a binary "Do Not Track" signal through to more complex expressions in cookie consent dialogues. Such signals are intended to influence how entities including websites and third parties process personal data. Web actors may collect privacy preferences in the hope of legitimizing data processing in the eyes of customers or to satisfy legal obligations.

Efforts to standardize privacy preferences go back to at least P3P, which was presented as a prototype to US regulators in 1997 and recommended as a standard by the World Wide Web Consortium (W3C) in 2002. It was adopted by around 20k websites [1], but was criticized by privacy advocates for not establishing consequences for false reporting of privacy practices [2]. Another W3C working group was formed in 2011 to specify the Do Not Track HTTP extension but it was closed before completion, citing the lack of planned support among "the ecosystem at large" [3] as exemplified by the Interactive Advertising Bureau's withdrawal [4]. The first wave of privacy preference signals is completed by the opt-out cookies [5] created by the Network Advertising Initiative (NAI) as part of a regulatory compromise with the Federal Trade Commission [6]. The NAI never published a specification, the opt-out only concerned a narrow definition of tracking, and very few vendors participated [5].

A second wave of privacy preference signals was prompted by the passage of privacy laws like the EU General Data Protection Regulation (GDPR) and the California Consumer Privacy Act (CCPA). For example, the GDPR establishes that an opt-in consent signal may constitute a legal basis for processing personal data providing the consent was "freely given, specific, informed and unambiguous". These laws prompted research that has largely focused on the interfaces through which opt-in [7, 8, 9, 10] and opt-out [11, 12] signals are collected. An ecosystem of actors has emerged to manage the collection of opt-in consent signals on behalf of websites [13]. Often these signals are collected and shared with a pay-for-membership "Global Vendor List", which has been termed the "commodification of consent" [14].

At this point, skeptics will rightly state that such signals exist in the world of *soft privacy* with no technical guarantees about personal data flows and that we should instead focus on the technologies associated with *hard privacy*. Such skepticism is compelling but should be qualified by the behavior of privacy advocates and AdTech firms. Both sides invested resources in P3P and DNT working groups. The latter posed a threat to AdTech business models as evidenced by the Interactive Advertising Bureau withdrawing from the working group after Microsoft announced it would be turned on by default [4]. The power of these signals can also be seen in websites' dark patterns that nudge users towards expressing certain preferences [9, 11, 10]. Given the stakes have been further increased by sanctions associated with the GDPR and the

CCPA, widespread adoption of a privacy preference signal would have privacy implications.

In terms of technical design, there is disagreement over who controls the interface by which users set privacy preferences. In both P3P and DNT, the user expresses preferences to a user agent. In contrast, user preferences are collected by embedding an interface in a web page in both of the approaches developed by AdTech industry bodies, namely the Interactive Advertising Bureau (IAB) [8] and the Network Advertising Initiative (NAI) [5]. This bypasses browsers by making the signal backwards compatible with existing technology. Turning to semantics, AdTech vendors proposed opt-in signals that could represent compliance, whereas privacy advocates proposed (global) opt-out signals that empower users. In summary, these signals have a long history and also have privacy implications going forward.

This paper systematizes historical knowledge on privacy preference signals (the past), measures which signals have been adopted as of February 2021 (the present), and reflects on adoption strategies for a pro-privacy signal (the future). We show a grim state of affairs for user control over privacy: P3P is obsolete, NAI's system still has only 75 participating AdTech firms, and the reincarnation of Do Not Track—the Global Privacy Control—has been adopted by less than 10 websites. Meanwhile, the Interactive Advertising Bureau's TCF 1.x and TCF 2.0 have been adopted by thousands of websites. We then use high-frequency web measurements to build a longitudinal case-study of how adoption and TCF 2.0 migration varied over time, websites and AdTech vendors. Our contributions include:

- **Systematize knowledge** about first wave (P3P, DNT, and NAI opt-out) and second wave (TCF and GPC) privacy preference signals.
- **Measure present day adoption** and show that TCF adoption is roughly comparable to historical P3P adoption among websites, whereas an order of magnitude more AdTech vendors have adopted TCF than all other signals combined.
- **Test explanatory variables** for TCF adoption like website popularity, category, number of embedded third parties, and presence of Google Ads. TCF adoption is higher among websites with closer ties to AdTech.
- **Longitudinal case-study** exploring TCF 2.0 migration strategies among the two most popular Consent Management Platforms, and how the new version changed the legal basis that individual AdTech vendors claim for tracking.

Section 4.2 describes the five privacy preference signals and Section 4.3 identifies related work measuring their adoption. This motivates our empirical measurements, which are described in Section 4.4. Our results describing the present are contained in Section 4.5. Section 4.6 discusses the past, present and future of privacy preferences. We conclude in Section 4.7.

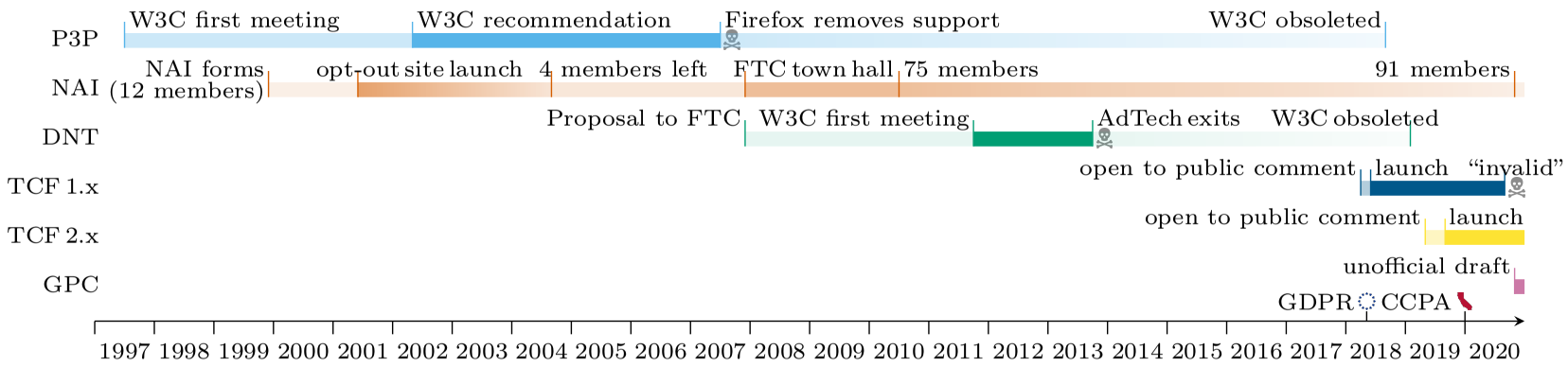


Figure 4.1: Timeline of key events for privacy preference signals and relevant laws

4.2 BACKGROUND

This section compares five privacy preference signals in terms of design properties and real-world adoption, which is summarized in Table 4.1. We selected these signals because they were the most widely adopted among the key stakeholders, namely browsers, AdTech vendors and websites. We do not provide a background on the widespread online tracking that motivate privacy preference signals, such as cookies [15, 16] and other tracking technologies [17, 18, 19]. Similarly, we do not consider privacy preserving technologies unless they function to express privacy preferences, such as when browsers/add-ons collect user preferences and automate sending the signal. We now turn to the five signals. Figure 4.1 provides and overview of the key events for each signal and Figure 4.2 provides a visual summary of the signal's flow.

4.2.1 *Platform for Privacy Preferences (P3P)*

P3P is one of the earliest privacy preference signals proposed for the Web. A demonstration of a P3P prototype was presented before the FTC in June 1997. The W3C recommended the P3P 1.0 specification in 2002, which describes an XML format to encode a human-readable privacy policy into a machine-readable specification stating the type, recipients and purposes of data collected. Users can define individual privacy preferences, which browsers can cross-check against a website's self-reported P3P policy. Each website's implementation could become arbitrarily complex with different policies for each web page and third-party cookie.

P3P was adopted by, respectively, 588 (10%), 463 (8.34%), 2.3k (2.3%), and 33.1k (60%) of the sites in samples from 2003 [20], 2007 [21], 2007 [22], and 2010 [1]. The final sample [1] is not representative of the wider web because the majority of sites were discovered by the Privacy Finder search engine, which specifically aimed to identify web sites that respect a user's privacy. However, the finding of 19 820 websites [1] implementing P3P in 2010 serves as a reasonable lower bound in Table 4.1. The same study [1] found that 11 (15%) of a sample of AdTech vendors had a P3P privacy policy.

Microsoft was the only browser developer to fully adopt P3P and stopped support in 2016. Mozilla supported only some P3P features, but removed them by 2007. Other browsers shunned P3P and instead allowed users to set defaults like blocking all third party cookies [23]. P3P-specific browser extensions provide a more meaningful perspective on conscious user adoption than usage statistics for each browser. For example, Privacy Bird, an add-on for Internet Explorer 5 and 6 that displays a website's P3P policy in an easy to understand language, was downloaded 20k times in the first 6 months [24].

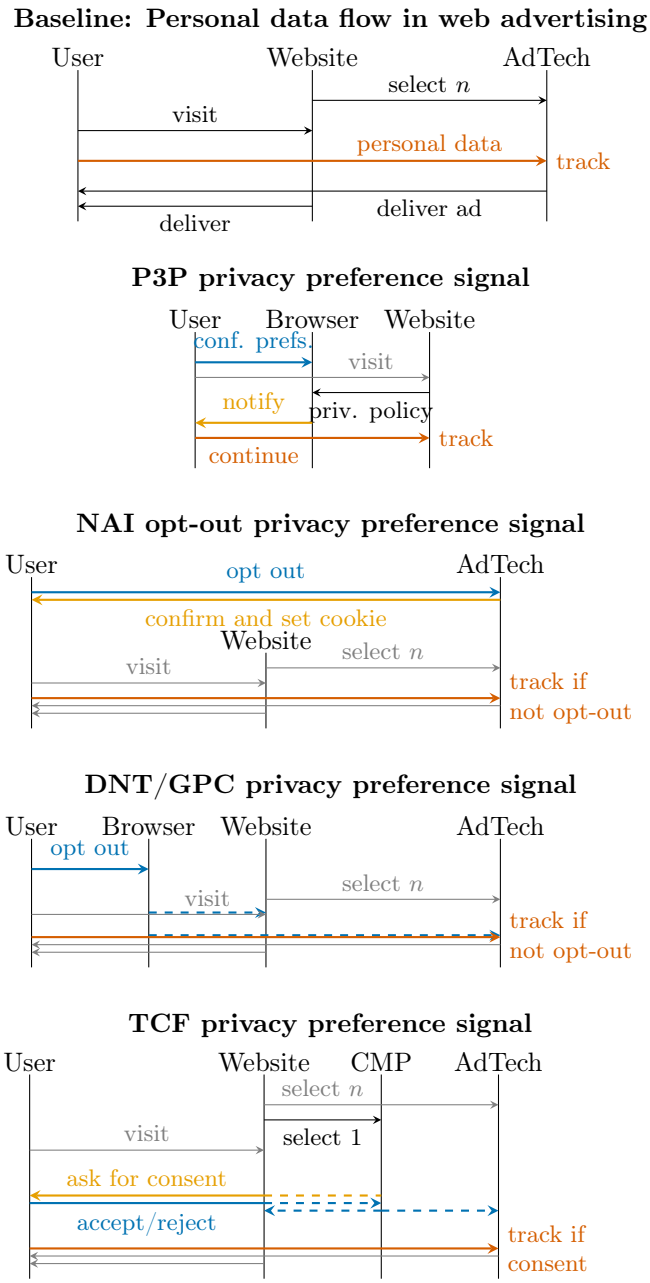


Figure 4.2: User prompt, privacy preference signals, and personal data flows when using each approach.

4.2.2 *Network Advertising Initiative (NAI) Opt-Out*

AdTech vendors founded a self-regulatory body, the NAI, as a compromise following the Federal Trade Commission's (FTC) report on web privacy submitted to Congress in 1998 [6]. The NAI established a system of opt-out cookies. Users can visit the NAI's website[1] and set an opt-out cookie for each participating vendor to signal that the user does not want to be tracked by

[1] https://optout.networkadvertising.org/

that firm. Critics [5] note that the NAI's narrow definition of tracking would not cover many techniques observed in the wild [17, 18, 19].

The NAI provide a list of all participating vendors, which was just 4 in 2004, 75 in 2010 [1] and stands at 75 participating vendors as of January 2021. Websites and browsers do not need to adopt the NAI's system because it piggy-backs on existing browser cookie functionality. The NAI reported one million visits to the the opt-out page in 2006 [5] but we cannot differentiate unique visitors. Returning to browser extensions, there were at least 44.9k users of the Targeted Advertising Cookie Opt-Out (TACO) add-on[2], which maintained an up to date list of opt-out cookies.

4.2.3 *Do Not Track (DNT)*

Acknowledging the failure of P3P, the W3C created a working group in 2011 to standardize the Do Not Track (DNT) mechanism [25]. DNT was less expressive than P3P. Implementation involved browsers sending a `DNT: 1` header with each HTTP request to signal that their user did not wish to be tracked. Stakeholders disagreed on whether DNT should default to on or off [4, 26]. This opposition was part of the reason why the W3C working group was closed without success in 2019 [3].

DNT was implemented in browsers by Microsoft, Apple, Mozilla and eventually Google [27]. Websites and third-party vendors could signal in an HTTP response header if they respected the user's DNT signal. This signal was not exposed in any browser's user interface[3] (outside of add-ons), which meant users were largely unaware of website adoption. Only 9 companies issued public statements regarding support of DNT [28]. In 2011, Mozilla reported DNT adoption by Firefox users to be at 17% in the US and 11% outside [29], although this oversamples privacy aware users.

4.2.4 *Global Privacy Control (GPC)*

The unofficial GPC draft specification [30], which was released in October 2020, continues the work of DNT in extending HTTP requests with a single bit value. Perhaps the most important change is re-framing *Do Not Track* as a "Do Not Sell" and "Object To Processing" signal, which is closer to the language of the GDPR and the CCPA, which became effective in May 2018 and January 2020. This means GPC references (enforceable) laws, which DNT lacked.

As of February 2021, Mozilla and the Brave browser are listed as publicly supporting GPC, but only Brave have implemented it. We do not provide any estimates for user size given it was released so recently.

[2] https://web.archive.org/web/20110920055245/https://addons.mozilla.org/en-us/firefox/addon/targeted-advertising-cookie-op/

[3] https://www.w3.org/TR/tracking-dnt/#responding

Table 4.1: Comparison of privacy preference signals

	P3P	NAI Opt-Out	DNT	GPC	Transparency & Consent Framework	
					TCF 1.x	TCF 2.x
General						
Convened by	W3C	AdTech & FTC	Privacy Advocates		AdTech	
Legal Basis	none	self-regulat.	self-regulat.	CCPA	GDPR	
Standardized by	W3C	NAI	W3C	GPC	IAB	
Design Properties						
Implementation	Privacy Policy XML	Opt-Out Cookie	HTTP Header		Consent Cookie from CMP	
User Interface	UA indicator	central website	UA setting or ext.		dialog on website	
User Decision	configure prefs.	opt out	turn on		select allowed purposes	
Decision Scope	all browsing	cookie lifetime	all browsing		website until re-request	
Vendor Decision	define policy	join NAI	adopt standard		declare processing purposes	
Website Decision	define policy	none	adopt standard		pick vendors	vendors + purposes
Adoption						
Websites	> 20k [20, 21, 1, 22]	–	> 9 [28]	?	> 1,539*	> 6,726*
AdTech Vendors	> 11 [1]	> 75 [1]	≈ 0 [28]	?	602	684
Browsers		–			–	–

? = unknown, – = compat. with exist. tech., * = in Tranco 100k, see Sec. 4.5, Safari Brave Chrome Internet Explorer Firefox

4.2.5 *Transparency and Consent Framework (TCF)*

After the enactment of the GDPR, an advertising industry body (IAB Europe) formed a working group to develop the Transparency and Consent Framework (TCF), "the only GDPR consent solution built by the industry for the industry" [31]. Participants predominantly representing private firms from the advertising and publishing industries co-developed the TCF, which defines the legal terms and data processing purposes that users consent to and the format by which consent signals are stored and exchanged between third parties. A new version (TCF 2.0) was introduced in 2020.

TCF is implemented by websites in the form of a consent dialog that does not require browser buy-in, much like NAI. It creates the role of *Consent Management Providers (CMPs)*, who implement the framework on individual websites. CMPs are central to the TCF in providing an interface between website, user, and ad vendors. They provide websites with a (customizable) cookie prompt to embed, store users' choices as browser cookies, and provide an API for advertisers to access this information. We refer to Hils et al. [13, Fig. 2] for a visual depiction of the ecosystem.

The IAB maintains a public list of CMPs, which lists 119 participating providers as of February 2021.[4] A website wishing to implement the TCF independently must become a CMP, otherwise they can out-source this to an existing CMP. In reality, a handful of CMPs dominate the market [8]. The largest CMPs are OneTrust and Quantcast, which account for 37.4% of all CMP implementations in the Tranco 100k (see Section 4.5).

To receive TCF consent signals from CMPs, AdTech vendors must register with the IAB and pay a yearly maintenance fee to join the *Global Vendor List (GVL)*[5]. As of Feb. 2021, 684 companies are registered on this list. Most CMPs collect consent for the entire GVL by default, which means privacy preferences apply to the whole list [14].

The specifications of TCF 1.x and TCF 2.0 both define a more complex signal than DNT/GPC. Under TCF 1.x, users may affirmatively consent to any combination of five data processing purposes. They may also state individual preferences for each vendor on the GVL. TCF 2.0 expands this model to ten purposes and two special features, increasing complexity even further.

In both TCF versions, users are prevented from expressing certain preferences. Vendors can claim that they have a legitimate interest in a specific purpose, which serves as their legal basis to process data even if the user clicks "Reject all". Starting with TCF 2.0, some CMPs provide users with the additional option to object to this processing (GDPR asks for such functionality), but this needs to be done separately in a subdialog. As such, the "Reject all" button commonly does not actually express *all* possible preferences. With TCF 2.x, vendors can declare that their legal basis is flexible. This means they would like to process data with the user's consent, but they can also perform

[4] https://iabeurope.eu/cmp-list/
[5] https://iabeurope.eu/vendor-list/

(limited) processing based on a legitimate interest. As the only exception, TCF 2.x removes the option for vendors to claim a legitimate interest in Purpose 1—"Store and/or access information on a device"—, possibly preempting an intervention by regulators. The policy changes between TCF 1.x and TCF 2.0 motivate measuring the transition.

4.3 RELATED WORK

Section 4.3.1 briefly describes the privacy practices employed by websites in order to motivate why privacy preferences matter. Section 4.3.2 surveys research into privacy preferences including the previous five signals. Section 4.3.3 links the paper to the general question of *why are technical standards adopted?*

4.3.1 *Privacy Practices*

Researchers consistently demonstrate privacy eroding techniques deployed in the wild [15, 16, 17, 18, 19] motivated by online advertising business models [32]. Personal data is leaked via social networks [33], third-party web scripts [34], apps [35], software development kits [36], and organizational breaches [37]. The scale of tracking motivate re-designing systems to provide privacy guarantees. For example, multihoming can be used to defend against fingerprinting [38] and trusted hardware can ensure compliance to stated privacy policies [39].

Turning to so-called soft privacy, data processors are constrained by law and social norms. These constraints are far from absolute. For example, half of websites in a 2017 sample violated laws implementing the EU Privacy Directive by installing cookies before collecting user consent [40]. This is likely because organizations do not incur significant costs following data breaches and privacy violations in terms of either regulatory fines or lost shareholder value [41]. Nevertheless, firms' privacy practices are *somewhat* impacted by data processors' self-declared privacy policies [42, 43, 44, 45] and even the privacy preferences expressed by users, to which we now turn.

4.3.2 *Privacy Preferences*

Interviews [46] and surveys [47, 48] can use natural language to understand users' actual privacy preferences, which tend to contradict observed behavior [49, 50, 51]. Privacy languages aim to express preferences more precisely than natural language. For example, APPEL encodes user preferences to be compared against P3P policies [52]. It could not express acceptable practices nor capture the realities of secondary sharing, which motivated XPref [53] and P2U [54], respectively. Alternative languages focus on the usability for developers [55], enabling audits [56], and providing explanations [57]. Privacy languages have been regularly surveyed by academics [58, 59, 60, 61] but

unfortunately there has been little adoption in practice [59]. This motivates our focus on signals deployed in the Web ecosystem.

In terms of the first wave of signals, measurements of DNT and NAI opt-out adoption relied on organizations disclosing private data sources like Firefox configurations [29], opt-out web page visits [5], or the NAI's membership [5]. P3P differed in that website adoption could be quantified via web scraping [20, 21, 62, 1, 22, 63] often sampling via commercial website rankings.

Turning to the second wave, there are no GPC adoption studies because only a draft specification has been released so far. The TCF ecosystem has been probed from a range of academic disciplines. Legal methods are relevant to the semantic content of the signal. For example, the purposes for collecting personal data standardized in the TCF may not be specific enough [64].

User interface research is important because the TCF does not standardize how the consent decision is presented to users, which is known to be influential [65, 66, 67, 10]. At least two studies have found that consent dialogues used to collect consent under the TCF contain design choices that nudge users towards providing consent [7, 9].

Web scraping studies have focused on implementation problems with TCF [64] or the ecosystem of consent management providers (CMP) [13]. These studies provide measurements of TCF in passing. For example, both studies measure TCF vendor registrations and their claimed purposes for processing data for TCF 1.x [13, p. 9] and both TCF 1.x and TCF 2.0 [64]. The latter study measures aggregate TCF 2.0 adoption, whereas we measure and visualise at the vendor level. Matte et al. [8] show how TCF 1.x adoption varies by top-level domain (TLD) and identify the most popular CMPs across the top 1k sites in five EU country code TLDs. Hils et al. [13] use longitudinal measurements to show the market growth of six CMPs, highlighting how fast the ecosystem changes.

4.3.3 *Standards Adoption*

We build on a body of work emphasizing the role of institutions in technical standards adoption. For example, many vendors initially saw the TCP/IP protocols as a nuisance [68]. Leiner et al. [68] describe how a series of "conferences, tutorials, design meetings and workshops" were organized to educate a generation of vendors and engineers. The rest is history.

The community was slow to turn to adoption questions like "What Makes for a Successful Protocol?", which was posed by RFC 5218 in 2008. Noting the qualitative nature of the resulting research, Nikkah et al. [69] provide an illuminating statistical analysis of the association between technical features of 250 RFCs and adoption success. Analysing unchanging technical features cannot explain why it took two decades before IPv6 was widely adopted [70, 71]. Economic considerations like the scarcity of IPv4 addresses and the supply of compatible hardware can help explain *when* standards are adopted [72].

Thus, standards should be considered in the context of wider ecosystems governed by economic incentives. For example, HTTPS adoption relies on X.509 certificate infrastructure that was "in a sorry state" in 2011 with many websites relying on shared or invalid certificates [73]. The situation was worse in the long tail likely because certificates are costly [74]. Felt et al. [75] report on significant improvements in 2017 and attribute improvements in the long tail to institutions like Let's Encrypt and publishing platforms—we show how similar economic considerations explain why TCF was adopted.

4.3.4 *Contribution*

Our main empirical contribution involves measuring the adoption of privacy preference signals among websites as of February 2021. Following the demise of P3P and DNT, the TCF has become dominant and the Global Privacy Control is still in its infancy. We explore variables explaining which websites adopt TCF, and also longitudinally measure migration to a new version (TCF 2.0).

This work differs from existing work by focusing exclusively on the adoption of privacy preference signals. We largely ignore the actors [13, 64] and interfaces [7, 9, 10] harvesting such signals and instead focus on which factors (e.g. website type, popularity, and partners) are associated with TCF adoption. Further, we are the first to systematize strands of research ranging from works in the late 1990s to post-GDPR studies. Finally, we provide the first results about migrating between versions of such signals using our the longitudinal methodology introduced in [13]. Our previous work focuses on detecting specific CMPs, some of whom collect non-TCF signals exclusively or only collect TCF signals for a subset of customers.

4.4 METHODS

We adopt a mixed approach[6] conducting both longitudinal high-frequency measurements to determine historic adoption of TCF and migration between versions, as well as a large-scale snapshot measurement to examine site-specific factors that may influence adoption. Section 4.4.1 describes our snapshot measurement of the Tranco 100k toplist. Section 4.4.2 explains how we use the Netograph platform to conduct longitudinal high-frequency measurements.

4.4.1 *Snapshot Measurements*

To measure the prevalence of TCF and its different versions on the web, we crawled the top 100k entries from the Tranco toplist, which aggregates the ranks from the lists provided by Alexa, Cisco Umbrella, Majestic, and Quantcast [76]. Our automated browser crawls were performed in February

[6] **Supplementary Material:**
https://github.com/mhils/pets2021-privacy-preference-signals

Table 4.2: Data sources for figures.

Figure	Approach	Data Source	N	CMP
4.3, 4.5	Snapshot (Feb. '21)	Tranco Toplist	100k	all
4.4	Snapshot (Feb. '21)	Tranco Toplist	10k	all
4.6, 4.9	Longitudinal	Netograph	7.2M	QC/OT
4.7	Longitudinal	Netograph	5.7M	QC
4.8	Longitudinal	Netograph	1.4M	OT
4.10–4.11	Diff. of vendor list	IAB	293	–

2021 using a Tranco toplist from January 2020[7]. We used this older toplist dated shortly before publishers transitioned to TCF 2.x in order to avoid survivorship bias in our observations. Picking a later toplist would over-sample websites created post-2020 who are certain to adopt TCF 2.0 and de facto avoid a migration decision. Our toplist and a current Tranco toplist (Tranco id KGNW from Feb. 19th 2021) overlap by 76.5%.

We first converted the Tranco list of domains to a list of URLs that can be crawled. For each *domain*, we attempted to establish a TLS and a TCP connection with www.*domain* and *domain* on port 443 and 80, respectively. This was repeated three times over a week to catch temporary service disruptions. We then picked a configuration that was reachable at least once, preferring TLS over TCP and secondly www.*domain* over *domain* to construct our crawl URL. An error in the TLS certificate verification was treated as unreachable. We used *http://domain* as a fallback if no connections were successful.

Our crawling infrastructure was set up in a European university network. Websites were opened using Google Chrome on Linux with its current default user agent,[6] a desktop resolution of 1024×800, and `en-US` as the preferred browser language. All other settings were set to their defaults: third party cookies are allowed, the DNT and GPC HTTP headers are not set. The low desktop resolution and all other settings were chosen to match that of our longitudinal measurements described below. Crawls are automated using custom browser instrumentation based on the Chrome DevTools Protocol. Unsuccessful crawls were retried twice within a week.

For every capture, we collected the following data points using custom browser instrumentation. First, HTTP headers are stored for all requests and responses. Second, connection-related metadata such as IP addresses and TLS certificate chains are logged. Third, for every domain in a capture, its relation to the main page, all cookies, IndexedDB, LocalStorage, SessionStorage and WebSQL records are saved. Fourth, we store the browser's DOM tree and record a full-page screenshot (including scrolling).

TCF Adoption

We automatically detect whether crawled websites implement the TCF. To do this, we wait for the website's DOMContentLoaded event to fire, then wait another ten seconds, and then inject JavaScript code into the execution

[7] Available at https://tranco-list.eu/list/K8JW

context of the root document. This approach for CMP detection was already validated by Matte et al. [8] with more aggressive timeouts. As each CMP must implement a `__cmp()` function for TCF 1.x and `__tcfapi()` function for TCF 2.x, we check for the presence of these functions to determine if TCF is being used. We additionally checked for other signs of TCF (such as the presence of `__tcfapiLocator` or `__cmpLocator`), but this search did not turn up any new results. For every TCF API we find, we issue a `ping` command to learn more about the implementation. In the case of TCF 2.x, the `PingReturn` object (as specified by the TCF) is expected to contain the CMP's identifier (as assigned by the IAB) as well as the CMP/GVL/TCF versions in use. We also considered that a CMP may masquerade as a different CMP here. We correlated the reported CMP ids with contacted domains and did not find any evidence of misrepresentation.

The adoption of TCF is naturally higher on some types of websites, such as those who typically display paid advertisements. To quantify this, we divided the Tranco 10k toplist into categories with the help of Symantec Rulespace [77], a categorization database already used in related work by Sanchez-Rola et al. [78]. We limit our analysis to the Tranco 10k as a non-negligible share of websites (11.7%) in the top 100k is not categorized, compared to only 2.4% for the top 10k websites. We note that recent work has shown that most categorization services are not fit for detecting specialized content or content-blocking [79], but this does not significantly affect our coarse classification of popular domains.

To determine the number of third parties present on each website, we normalized all requested URLs to their effective second-level domain using Mozilla's Public Suffix List [80]. This list contains all suffixes under which internet users can directly register names, including non-standard "TLDs" such as `blogspot.com`. We note that this approach does not account for recent obfuscation techniques such as CNAME cloaking [81].

We also examined the fraction of websites that appear to be collecting data versus those showing a cookie prompt. To determine a lower bound, we took all third-party domains that were included on at least 1.000 websites in the Tranco 100k (158 domains) and manually removed shared resources such as content delivery networks which may not constitute tracking (12 domains). We then determined for each website if any of the remaining 146 third parties were embedded. For example, we exclude `s3.amazonaws.com` as this domain is commonly used to serve static assets and not for tracking. In contrast, almost all remaining domains clearly belong to ad companies. We include both lists in the supplementary material.[6]

Finally, we estimated the prevalence of non-TCF cookie notices or consent prompts in our snapshot measurements using a simple back-of-the-envelope heuristic. For every capture, we scan the stored copy of the browser's final DOM tree for the occurrence of the phrase "cookie". The resulting estimates only indicate orders of magnitude, which is acceptable given they are not core to any of our results. Rather they are intended to provide context, such as

showing government websites are significantly less likely to present a cookie notice than our other categorizations (see Figure 4.4). In a manual inspection of 50 randomly picked domains with and 50 domains without "cookie" in their DOM tree, we found five domains that had a "Cookie Notice" link in their footer (but no dialog) and no false negatives (which yields a 5% error rate overall). Again, this part of our analysis is not as rigorous as our other measurements and is only intended to provide context in Figure 4.4.

4.4.2 *Longitudinal Measurements*

To measure the adoption and transition between TCF versions longitudinally, we analyze automated browser crawls recorded by the Netograph web measurement platform.[8] Netograph continuously ingests a live feed of social media posts, extracts all URLs, and visits them from crawlers located in EU and US data centers. For brevity, we refer to [13] for a discussion of the validity and reliability of this measurement method. Most importantly, HTTP message contents are not retained due to storage constraints, but a large amount of metadata is stored, such as the HTTP headers of every request.

Relying on metadata in our longitudinal data means we have to measure TCF adoption using CMP-specific indicators. Instead of building quick and dirty heuristics for over 90 CMPs, we focus our efforts on creating a set of reliable indicators for two of the leading providers in the consent management market, Quantcast and OneTrust, which are embedded on 9.7% of websites in the Tranco 10k (Feb. 2021). We manually analyzed their respective dialog implementations and identified distinct HTTP requests that indicate the use of specific TCF versions[6]. For Quantcast, we detected the use of TCF for all implementations dating back to May 2018. For OneTrust, we identified the use of TCF 1.x or TCF 2.0 in their Cookie Consent SDK launched at the end of 2019 (`otSDKStub.js`).

From Netograph's 177 million captures in the social media dataset, we obtained all 5.7 million captures that include a Quantcast consent dialog and all 1.4 million captures that include a OneTrust consent dialog. We grouped captures by their effective second-level domain to not overcount repeated measurements with varying subdomains. Due to Netograph's sampling strategy, less popular domains may not be observed for a several days. We account for this by explicitly marking the period between the last TCF 1.x and the first TCF 2.0 measurement as an (unobserved) transition phase.

Measuring Vendor Adoption

To track the adoption of TCF 2.0 by AdTech vendors, we downloaded all previously published lists of vendors registered as participating in the TCF from the IAB and verified their accuracy using the Internet Wayback Machine. These lists include each vendor's declared purposes for processing personal data.

[8] https://netograph.io/

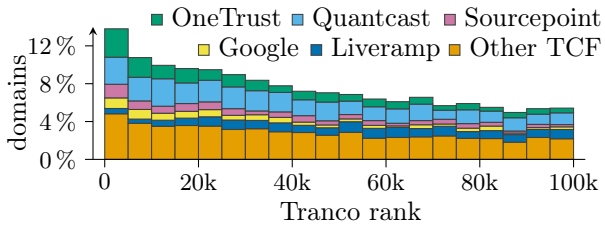


Figure 4.3: Share of websites in the Tranco 100k that use a CMP. OneTrust and Quantcast are the most popular providers, followed by Sourcepoint, Google, and Liveramp.

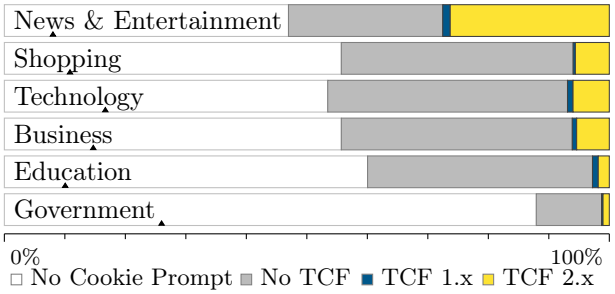


Figure 4.4: Share of websites in the Tranco 10k with a (TCF) cookie prompt. For reference, ▲ marks the share of websites which do not embed popular third parties.

As of Feb. 2021, there are 215 revisions of this list for TCF 1.x and 78 revisions for TCF 2.0. We then inspected these previous versions for longitudinal changes and measured every instance when an AdTech vendor joins, leaves, or switches to TCF 2.0. While TCF 2.0 is not backwards compatible from a publisher's point of view, a vendor that has declared support for TCF 2.0 may still accept TCF 1.x consent strings from publishers.

4.5 RESULTS

Section 4.5.1 focuses on the relationship between website characteristics and TCF adoption mainly using snapshot measurements. Section 4.5.2 explores how vendors and websites migrated to TCF 2.0 using our longitudinal approach. Table 4.2 maps each figure to the approach, data source, and covered CMPs. We provide the underlying data in the supplementary material.[6]

4.5.1 *TCF Adoption*

We first explore how TCF adoption varies by the popularity and category of website. Figure 4.3 shows that TCF is more prevalent among popular websites (e.g the Tranco 5k) and that adoption is relatively consistent through the Tranco 100k. Websites embedding OneTrust comprise a greater fraction of TCF implementations for more popular sites (Tranco 20k), whereas Quantcast embeds are more evenly distributed. Quantcast's free self-service solution may be better suited to less popular sites than OneTrust's, which requires

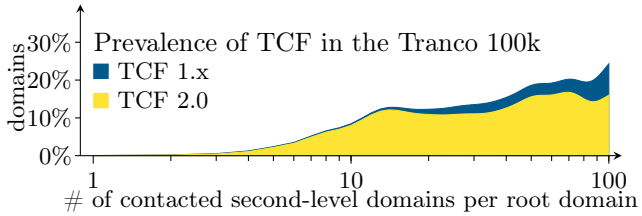


Figure 4.5: Adoption of the TCF increases significantly for websites that embed a large number of third parties.

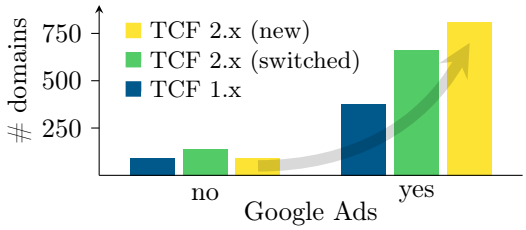


Figure 4.6: Google did not participate in TCF 1.x and only joined TCF 2.0. Their partners' websites were far more likely to adopt TCF 2.x but not TCF 1.x.

an interaction with a sales associate. By offering a free and usable solution, Quantcast is playing a similar role to Let's Encrypt with HTTPS adoption [75].

Figure 4.4 shows that TCF adoption in the Tranco top 10k is highest among websites classified as News & Entertainment and is lowest among Government websites. The grey bars provide a relatively coarse indication (see the previous section) of what percentage of each category displays a cookie prompts. Few Government websites display prompts, which helps to explain the low TCF adoption. Almost half of all cookie prompts on News & Entertainment sites implement TCF, whereas this fraction is less than 15% for each of the other five even though the first five categories have a similar fraction of websites showing cookie prompts. This motivates exploring alternative explanations.

We explored whether web relationships can help explain varying adoption rates. Figure 4.5 shows that TCF adoption increases with the number of embedded third parties. This result could be caused by third parties influencing partner websites to adopt TCF, but it could also be mere correlation. Websites with business models based on personal data may be *both* more likely to embed many third parties and also more likely to adopt the TCF.

Causality could be probed via a natural experiment in which websites were randomly assigned a partner that exerts influence. It can be argued the decision of Google to join TCF 2.0 but not TCF 1.x provides such an opportunity. By comparing the relative adoption of TCF 1.x and TCF 2.0 among websites which embed Google with those who do not, we can isolate the effect on TCF adoption of partnering with Google. If partnering with Google influences websites' decisions, we would expect a higher fraction of such websites to adopt TCF 2.0 but not TCF 1.x as compared to the same fraction among non-partners. Indeed, Figure 4.6 shows that for websites supporting TCF 2.x and not using Google Ads, 60% had already joined TCF 1.x, whereas

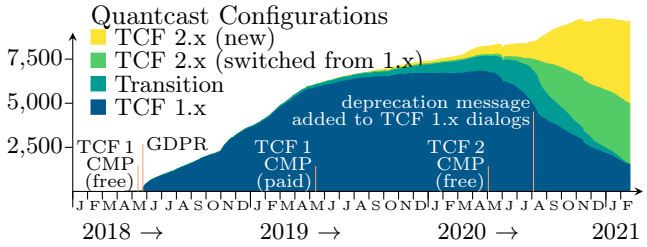


Figure 4.7: TCF Adoption by Quantcast customers. Note that the y-axis differs from OneTrust; Quantcast started with a significantly larger number of TCF 1.x customers.

this applies to only 45% of the websites using Google Ads. We cannot tell whether the influence is active (e.g. vendor X only contracts with TCF websites) or passive (e.g. website Y finds it easier to adopt the same standard as their partners).

To shed more light on these relationships, we run logistic regressions with TCF 2.0 adoption as the dependent variable. For each website, we have the following explanatory variables: a binary dummy for the presence of Google ads β_1 (from Figure 4.6), log of the number of embedded third parties β_2[9] (from Figure 4.5), and the website category (from Figure 4.4). We include a full regression table in the Appendix (Table 4.3).

As we would expect from the figures, the first regression shows β_1 and β_2 have a positive relationship with adoption:

$$y \approx -4.6^{***} + 0.15^{***}\beta_1 + 0.77^{***}\beta_2 \tag{1}$$

and both effects are statistically significant at the $p = 0.01$ level. This means each variable adds additional explanatory power.

Model 2 adds a fixed effect for each website category and this boosts the Pseduo-R^2 from 0.08 to 0.13 relative to Model 1. The coefficient for News & Entertainment is positive and highly significant. The high adoption rate among such websites exceeds what could be explained by β_1 and β_2 alone.

Finally, Model 3 explores the interaction effect between β_1 and β_2. The sign of $\beta_1 * \beta_2$ means that the relationships are sub-additive—the increased likelihood of adoption from increasing both variables is less than the sum of increasing each variable independently. Although these regressions have shown that website category and web relationships help explain TCF 2.0 adoption rates, the Pseduo-R^2 shows a lot of the variance remains unexplained. This could be down to our relatively crude statistical design aiming to directly link variables to organisation-level outcomes. A recent systematization of knowledge [41] highlights similar difficulties explaining cybersecurity outcomes via manifest variables and suggests latent variables inferred via reflexive indicators represent a better way forward.

[9] We count the first party domain so that $\beta_2 \geq 0$.

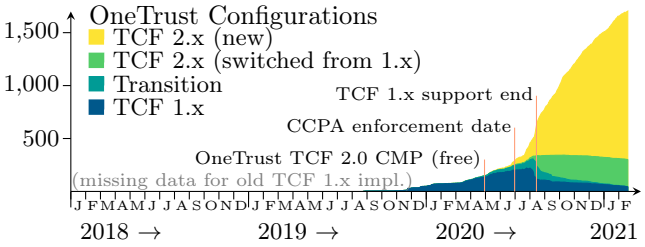


Figure 4.8: TCF Adoption by OneTrust customers. Most TCF 1.x customers switched to TCF 2.x around August 2020. Since July 2020, OneTrust gained a large number of new customers which directly started using TCF 2.x. *Transition* marks the unobserved interval during which a switch from TCF 1.x to 2.x occurred.

Figure 4.9: Share of websites in each segment of the Tranco toplist that use the TCF and have upgraded to version 2.x.

4.5.2 *TCF 2.0 Migration*

The release of TCF 2.0 provides an opportunity to observe how actively both vendors and websites adopt these signals.

Websites

Quantcast have the most customers embedding TCF, claim to be a driving force behind its development, and launched a new free TCF 2.0 product in May 2020. Yet Figure 4.7 shows how a large share of their customers had not adopted the new version when TCF 1.x support by the IAB ended on August 15th. Approaching the IAB's deadline, Quantcast went as far as embedding a prominent deprecation notice visible to all website visitors into its TCF 1.x consent dialogs (see Figure 4.13). Quantcast lost customers while enforcing the switch over, which can be seen in the fall (6%) in old customers who had implemented TCF 1.x from the start of August to end of September. Quantcast's total customers continue to grow due to new customers who directly adopt TCF 2.0 (the yellow fraction), but the fall in old customers can be seen in the decreasing total of the green and blue lines in Figure 4.7.

In contrast, OneTrust lost very few customers in transition, which can be seen in the bright green area in Figure 4.8. OneTrust acquired many new customers from June 2020 and the majority of these immediately adopted TCF 2.0. As a result, OneTrust had a higher fraction of customer implementing TCF 2.0 than Quantcast by the end of September 2020 even though Quantcast pursued a more assertive transition strategy. However, Quantcast remain

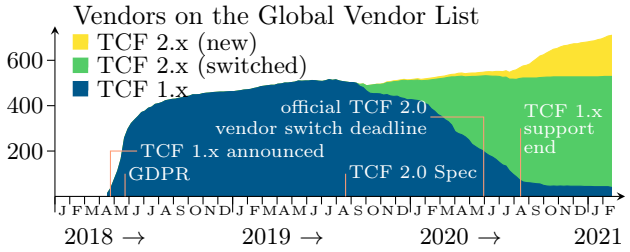


Figure 4.10: TCF Adoption by ad-tech vendors.

comfortably ahead of OneTrust in terms of number of websites embedding TCF (although OneTrust also implements a significant number of non-TCF dialogs [13]).

Returning to the role of top list position, Figure 4.9 shows that websites in the Tranco top 100 began experimenting with TCF 2.0 migration in the first half of 2020. The experimentation can be seen in how migration went down at various points. The majority had permanently transitioned by July 2020. This suggests the CMP's announcement about ending support for TCF 1.x were sufficient to lead to migration for popular websites. However, the less popular websites were far less responsive.

Vendors

The majority of early adopters were vendors rather than websites. By the start of 2020, more vendors (84) had switched to TCF 2.0 than there were websites (48) embedding either version of TCF using OneTrust's Consent SDK. Figure 4.10 shows vendors appear to follow an S-growth pattern with slow uptake, a relatively small window in which the majority adopt, and a stubborn tail. The number of vendors implementing each version of TCF was relatively consistent through to September 2020, which suggests the upgraded TCF was not a major draw for vendors unlike for websites embedding Google Ads (see Figure 4.6). The growth rate increased from September 2020 for reasons we do not know, but this is much smaller than the post-GDPR growth.

Comparing time to adoption and migration between vendors and websites speaks to the question of which constituency is driving TCF adoption. Figure 4.10 shows most vendors had already adopted TCF 1.x by the time GDPR came into effect, whereas OneTrust had no TCF product and only a fraction of Quantcast's 2020 customers were implementing TCF. The same pattern holds for TCF 2.0 migration. This is consistent with vendors providing an incentive for partner websites towards adoption. While we cannot claim causality, this evidence at least makes it unlikely that websites pushed vendors towards adoption.

Implications

Thus far we have focused on adoption and migration without considering the details or privacy implications of the switch. We illustrate the need for future

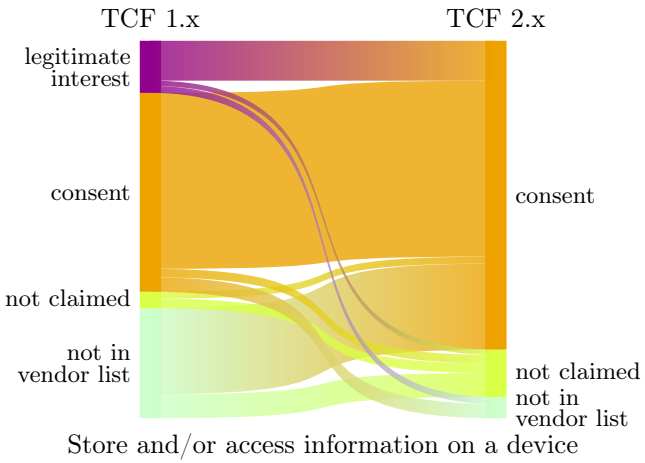


Figure 4.11: Removing the option to claim legitimate interest for purpose 1 of the TCF (see Section 4.2) led more vendors to collect consent for accessing information such as advertising identifiers under TCF 2.x. New vendors that did not adopt TCF 1.x (*not in vendor list*) mostly seek consent as well.

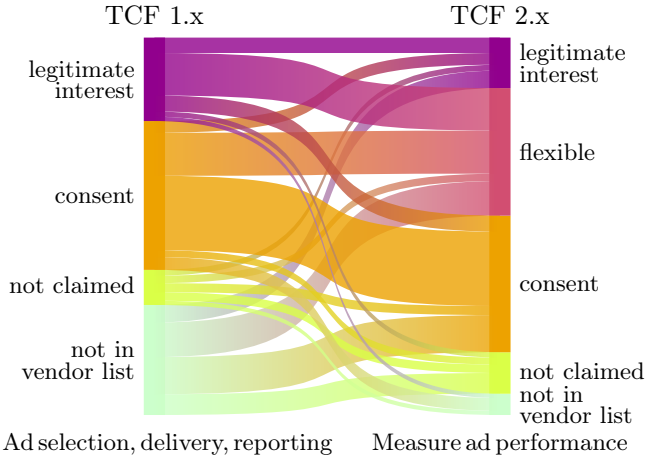


Figure 4.12: In migrating from TCF 1.x to TCF 2.x, a large portion of vendors now can claim to be flexible regarding the legal basis; i.e. they will perform the processing based on consent or a legitimate interest.

work by measuring the effect of migrating to TCF 2.0 on the legal basis by which vendors claimed the right to process personal data. We recount some of the background from Section 4.2. Both versions of TCF define purposes for processing personal data. For each purpose, vendors implementing TCF 1.x can declare either; they do not use personal data for that purpose, need to first obtain consent before doing so, or claim they have a legitimate interest in doing so (which users cannot dispute).

The IAB removed the option to claim a legitimate interest in storing and/or accessing information on a device under TCF 2.x. Figure 4.11 shows how this shifted the majority of vendors who were previously claiming legitimate interest towards asking for consent. This highlights how standards setters can influence how privacy preferences are communicated at scale by removing the legally questionable options.

Updated standards can also add complexity that makes analyzing impacts difficult to evaluate. For example, the purpose "ad selection, delivery and reporting" was renamed and split into multiple purposes in TCF 2.x. Additionally, vendors had the additional option to declare that they are flexible regarding the legal basis; they can perform the processing based on consent or a legitimate interest. Figure 4.12 shows how this led to a decrease in both the number claiming legitimate interest and also the number collecting consent, which means its unclear whether users lost or gained control under the new standard. These results show just one way in which the design of standards impacts user privacy.

4.6 DISCUSSION

This section discusses the past, present (as established in the previous section) and future of privacy preference signals.

4.6.1 *Past*

Mark Twain's quip that "history doesn't repeat itself, but it often rhymes" is also true of privacy preference signals, and identifying these rhymes helps to reason about the present and future. For example, Table 4.1 shows that signals proposed by AdTech (NAI and TCF) collect user preferences via a web page, whereas the signals proposed by privacy advocates are collected by a browser. As a result, browsers immediately support AdTech signals and could only stop them by actively preventing web content rendering, meanwhile AdTech vendors must actively make the decision to support P3P, DNT and GPC. Consequently, standards developed by AdTech industry bodies have been adopted by browsers by default, whereas AdTech vendors can delay adoption and thus undermine the standard.

Privacy preference signals also vary in terms of the signal's scope, permanence, and how decision volume scales with web usage. Table 4.1 highlights how privacy preferences are collected in a single interaction under P3P and DNT/GPC and the browser assumes that this decision applies to the entire Web. Consequently, the user makes a single decision that has long-term signaling implications. In contrast, the NAI's opt-out cookies only apply to specific forms of tracking [5] and only last until the user loses the cookie or the vendor sets a new one.

Scope and permanence are even narrower under the TCF, which contains asymmetries based on the preferences expressed. The decision not to provide consent[10] only applies to a specific website and only last until the website re-requests consent, whereas positive consent signals may apply to multiple websites [14, 8] and re-requests are less frequent. Table 4.1 shows history repeating itself in that privacy advocates support a signal that imposes a low

[10] Notably, the TCF framework does not even mention the possibility a user can "revoke" a decision [31].

decision load on users (P3P, DNT and GPC), whereas AdTech vendors support impermanent signals with a narrow application that force a decision burden on users (NAI and TCF).

Turning to the forum in which signals were designed, we have seen a movement away from development via consensus-based working groups committed to open standards. Initially all parties met in working groups coordinated by the W3C but the clashing political objectives led to splintering. For example, the Digital Advertising Alliance withdrew from the DNT working group in 2012 citing the lack of progress [82].

The second wave of privacy preference signals were developed outside of open, consensus-based groups. TCF was developed via a working group listing 139 participating organizations [83] for which the Interactive Advertising Bureau controlled membership. The resulting TCF signal is closed in that both websites and vendors need the IAB's permission to implement it, although this authority is delegated to consent management providers. GPC is developed more openly, but lists only 17 supporting organizations with no formal forum to coordinate development. For comparison, the P3P 1.0 specification lists participants from 56 organizations, the DNT working group contained 110 members [82], and the NAI for a long time only included "a fraction of the industry" [5] and now counts 91 members.

In retreating to less consensus-based processes, the Global Privacy Control and the Interactive Advertising Bureau follow (in more than just initials) the governance model of the Internet Advisory Board, which was created in 1984 to incorporate stakeholders beyond Vint Cerf's "kitchen cabinet" [84, p. 51]:

> "The IAB cannot be characterized as a democracy, since nobody voted and the Board only let in the people they wanted ... Democracy, with its competing factions and its political compromises, was not an appropriate political model for the IAB or the Internet."

The same could be true of privacy standards given over 10 years was spent drafting P3P and DNT at the W3C. It should be noted that the Internet's IAB later moved towards more open governance by creating and transferring power to the IETF [84]. It seems unlikely AdTech's IAB will voluntarily follow suit, which raises the question of regulatory involvement.

The history of privacy preference signals is intertwined with regulation. Do Not Track began as a letter to congress and was re-invigorated by the FTC chairman going off script to mention it years later [85]. The NAI's opt-out cookies resulted from an agreement with the FTC to self-regulate [5]. The IAB created the TCF in response to the GDPR, and GPC quotes "Do Not Sell" directly from the CCPA. However, none of these signals are mandated by law, which means they could become de-facto standards by achieving widespread adoption.

A final lesson from history is that for all the willingness of browser developers to attend working groups, they are reluctant to support privacy preference signals if doing so risks impacting user experience. For example, Microsoft

set allow-all cookies as the default for sites who misconfigure P3P presumably because blocking cookies may have affected those websites. This decision on defaults was widely exploited; a misconfiguration described on a Microsoft support page was detected down to the exact typo in 2 756 sites [1]. Similarly, DNT was adopted without sufficient enforcement from browsers, which does little to improve user privacy beyond shifting the blame to AdTech vendors for not respecting the signal.

More encouragingly, history also shows privacy advocates can subvert systems with relatively low-effort browser add-ons. For example, advertising networks expected every user to visit their individual websites to set opt-out cookies [5]. In reality, the TACO browser extension allowed one individual to maintain and share an updated list of cookies with thousands of users [85]. Similarly, the Privacy Bird allegedly helped boost P3P adoption by directly making the user aware of websites' adoption decisions. These two examples point to the importance of designing privacy enhancing technologies that allow users to send low-effort privacy preference signals. This becomes especially urgent given the state of the present, to which we now turn.

4.6.2 *Present*

Having surveyed a history in which P3P and DNT were eventually deprecated and NAI membership remains at less than one hundred vendors, our measurements provide an updated picture as of February 2021. TCF is the dominant signal as the GPC was released as an unofficial draft in October 2020 and only six websites in the Tranco top 100k now implement it. Given signals must be adopted by both sender and recipient, we now discuss adoption among each stakeholder.

Websites are arguably the most important stakeholder for the success of TCF since only websites can collect consent signals [14]. We discovered 7,582 TCF implementations in the top 100k. A crude comparison can be drawn with a 2010 sample detecting 19.8k P3P implementations [1]. Turning to estimates that reference a toplist, TCF is more prevalent among both the top 5k (13%) and top 100k (7%) than historic P3P measurements (8% [21] and 2% [22] respectively). Such comparisons are limited by changes in the Web and also research methods; P3P adoption studies relied on commercial rankings, whereas we used a top list designed to be stable over time for research purposes. This should make our measurement more comparable to future work.

Turning to adoption among AdTech vendors, vendors were early adopters of TCF and also the first to migrate to TCF 2.0 (see Figure 4.9). By October 2020, more than 600 vendors had adopted TCF. For comparison, just 75 vendors were offering opt-out cookies in June 2010 of which only 11 were also implementing P3P [1]. Although AdTech vendors drafted the TCF specification, adoption was not inevitable given the NAI had no more than 6 full members from 2001–2007 [5]. Thus, TCF is the first privacy preference signal to achieve widespread adoption among AdTech vendors.

Our results also speak to why websites are adopting TCF. Numerous pieces of evidence suggest vendors incentivize partner websites to adopt TCF (see Figure 4.3, Figure 4.5 and especially Figure 4.6). An interesting comparision can be made with P3P. Websites embedding more third-party domains are more likely to adopt TCF but less likely to adopt P3P [62, p. 292]. This supports the common sense intuition that TCF was designed to perpetuate privacy eroding business models.

More generally, we provide evidence in support of the general finding that private firms deploy economic resources to ensure the adoption of standards [86]. Figure 4.3 shows how Quantcast's free consent management solution supported TCF adoption, particularly among less popular sites. The role of institutional support is crucial even to open standards, such as the organization of TCP/IP education events [68] and subsidization of free certificates via Let's Encrypt to support HTTPS adoption [75]. In terms of migrating to updated standards, we show how Quantcast boosted TCF 2.0 adoption by adding prominent deprecation messages into consent dialogs. Thus, Figure 4.7 suggests that IAB policy (TCF 1.x consent strings becoming invalid) led to Quantcast losing customers.

Finally, we can quantify the relative decision volume of users relative to vendors. Quantcast boast of processing 25 billion consent signals [87], whereas we observed just 2,103 changes in vendor purposes since 2018. This means users have made at least 11 million times more decisions than vendors since TCF was launched. At 3.2s per decision [13], this means users have spent at least 2,500 years since 2018 expressing their privacy preferences through Quantcast dialogs alone.

4.6.3 *Future*

Given this startling time investment in sending TCF signals, it is worth considering what the future holds for pro-privacy signals. Releasing the GPC specification in an unofficial draft [30] over two years after GDPR came into effect and ten months after CCPA provided TCF with a first-mover advantage. However, we have few concerns that privacy aware users will adopt the GPC in the future. Pro-privacy browsers like Firefox supported the design, additionally the Brave browser[11] and add-ons like Privacy Badger[12] already turn the GPC signal on by default.

We are less optimistic that the intended recipients, namely AdTech vendors, will adopt the GPC signal. Much like with DNT [4], AdTech vendors are likely to claim that on-as-default makes the signal meaningless. However, privacy advocates can now rely on privacy laws like the CCPA, which was not available when DNT was first adopted by browsers.

Fighting legal cases to establish a favorable precedent is a likely strategy. One of the GPC's participating organizations, Brave Browser, has already

[11] https://brave.com/global-privacy-control/
[12] https://www.eff.org/gpc-privacy-badger

lodged complaints under the GDPR against rival browsers [88], national regulators [89], and even the IAB Europe's website [90]. We anticipate similar actions under the CCPA, especially given California's attorney general tweeted about the GPC in January 2021[13]. Multiple publishers adopting the same standard and out-sourcing implementation to dominant CMPs creates the potential for auditing at scale [p. 10][13], as evidenced by an NGO's threat of automated complaints against publishers[14].

Regulatory interventions may begin to undermine the adoption of TCF. For example, the Danish regulator ruled that the Danish Meteorological Institute could not claim a legitimate interest in collecting personal data [91]. Possibly preempting such a ruling, the option to declare a legitimate interest in storing and/or accessing information on a device was removed in TCF 2.0 (see Figure 4.11). The case also ruled that opt-out must be as easy as opt-in. Many websites collecting TCF signals do not follow this ruling [9, 13]. The leading provider of TCF dialogs distances itself from ambiguity in privacy law [92] by making the design choice a configuration that websites select, with one CMP warning "with great customizability comes great responsibility" [13]. This indicates that AdTech vendors perceive liability risk related to TCF.

This discussion raises the question of what happens when two signals co-exist. Whereas standards usually have a definitive winner, such as DVD over DIVX or VHS over Betamax [93], GPC and TCF signals can be sent simultaneously because they are defined on different network layers (see Table 4.1). Encouragingly, one could imagine a future in which browsers exploit control over what is rendered to the user to block dialogs from loading, whereas AdTech cannot stop browsers from sending GPC headers as part of HTTP requests. Signals co-existing is more troublesome when it comes to interpretation. A TCF opt-in signal could be sent in an HTTP request with GPC opt-out headers. We leave it to legal scholars and future court cases to ponder which signal has priority.

Arguably this back and forth over privacy preference signals has been a distraction for over 20 years. Regardless of the adoption of privacy preference signals, there is little basis to trust that expressed preferences will be respected. In terms of what we can observe: vendors ignoring the DNT signal was public policy [4], P3P was intentionally misconfigured by websites [1], TCF consent signals misreport the user's expressed preferences [8], tracking remains ubiquitous in a post-GDPR world [78] and there is growing evidence firms use dark patterns to manipulate users' expressed preferences [94, 95, 96]. More fundamentally, there is no way of auditing whether AdTech vendors respect expressed signals.

[13] https://digiday.com/media/why-a-tweet-from-californias-ag-about-a-global-privacy-tool-has-companies-scrambling/

[14] https://noyb.eu/en/noyb-aims-end-cookie-banner-terror-and-issues-more-500-gdpr-complaints

4.7 CONCLUSION

Privacy preference signals must be adopted by both senders (users) and recipients (AdTech vendors) who have differing requirements. Vendors want to receive positive consent signals in order to comply with privacy laws, and prefer not to receive negative signals that undermine the vendor's business model. This reasoning helps to explain why hundreds of vendors adopted TCF [13, 64], which represents a historical anomaly given vendors reluctance to adopt P3P [1], DNT [28] and NAI opt-out cookies [5]. Our evidence that vendors were early adopters of TCF 2.0 (Figure 4.9) underlies the AdTech vendors' commitment to receiving these signals.

History reveals two approaches to collecting users' privacy preferences that are represented in the signal, namely via the user agent (as in P3P and DNT) or a webpage (as in NAI opt-out). As with the previous signal designed by AdTech [5], TCF collects user preferences via dialogs embedded in a web page but this requires adoption among websites. Our results show website adoption varies from 5% to 12% across sections of the Tranco top 100k (Figure 4.3) and is most prevalent among News & Entertainment websites (Figure 4.4). We also show that the presence of Google Ads (Figure 4.6) and the number of embedded parties (Figure 4.5) are both associated with greater TCF adoption rates.

Adoption is further supported by AdTech actors like Quantcast lowering the cost of adopting TCF by providing free dialogs marketed as compliant with GDPR (although legality has been called into question [9, 8]). The increase in adoption following May 2018, which can be seen in Figure 4.7, shows how AdTech capitalised on the passage of the GDPR. This means AdTech firms now not only draft the TCF, but also actively manage and configure it. This market power facilitated the swift transition to TCF 2.0 (see Figure 4.8 and Figure 4.7), which is remarkable when contrasted against the time to migrate to HTTPS [75] or IPv6 [72].

Thus, our measurements of the present reveal TCF is now the dominant privacy preference signal. Further, its adoption among *both* senders and recipients is a significant historical development (see Table 4.1). Adoption among recipients is unsurprising given the working group who designed TCF was controlled by the Interactive Advertising Bureau and contained no privacy advocates. However, websites appear to have sided with their business partners over users. Consequently, users are forced to send signals via time consuming dialogs. Our back-of-the-envelope calculation on p. 4.6.2 suggests over two thousand years of user time has been spent on sending TCF consent signals since 2018. All stakeholders should ask to what extent the TCF's fine-grained, site-by-site signal clarifying privacy preferences has materially changed how recipients process personal data? A second question is whether a revised signal would lead to better outcomes, or can the problems only be resolved by the technical constraints of *hard privacy*?

ACKNOWLEDGEMENTS

We would like to thank Aldo Cortesi for his continuous support and the generous access to the Netograph API and capturing technology. We thank Anelia Kurteva, Jérémie Bernard Glossi, Dennis Jackson, our shepherd Christo Wilson, and the other anonymous reviewers for many constructive comments. The second author is funded by the European Commission's call H2020-MSCA-IF-2019 under grant number 894700.

REFERENCES

[1] Pedro Giovanni Leon, Lorrie Faith Cranor, Aleecia M McDonald, and Robert McGuire. Token attempt: The misrepresentation of website privacy policies through the misuse of P3P compact policy tokens. In *ACM Workshop on Privacy in the Electronic Society*, pages 93–104, 2010.

[2] Electronic Privacy Information Center and Junkbusters. Pretty Poor Privacy: An Assessment of P3P and Internet Privacy. https://epic.org/reports/prettypoorprivacy.html, 2000.

[3] Tracking Protection Working Group. WG closed. https://github.com/w3c/dnt/commit/5d85d6c, 2019.

[4] Interactive Advertising Bureau. "Do Not Track" set to "On" by Default in Internet Explorer 10—IAB Response. https://www.iab.com/news/do-not-track-set-to-on-by-default-in-internet-explorer-10iab-response/, 2012.

[5] Pam Dixon. *The Network Advertising Initiative: Failing at Consumer Protection and at Self-Regulation.* World Privacy Forum, 2007. http://www.worldprivacyforum.org/wp-content/uploads/2007/11/WPF_NAI_report_Nov2_2007fs.pdf.

[6] Martha K. Landesberg, Toby Milgrom Levin, Caroline G. Curtin, and Ori Lev. *Privacy online: A Report to Congress.* US Federal Trade Commission, 1998.

[7] Christine Utz, Martin Degeling, Sascha Fahl, Florian Schaub, and Thorsten Holz. (Un)informed Consent: Studying GDPR Consent Notices in the Field. In *Proceedings of the 2019 ACM SIGSAC Conference on Computer and Communications Security*, CCS '19, pages 973–990. ACM, 2019.

[8] Célestin Matte, Nataliia Bielova, and Cristiana Santos. Do Cookie Banners Respect my Choice? Measuring Legal Compliance of Banners from IAB Europe's Transparency and Consent Framework. In *IEEE Symposium on Security and Privacy*, pages 791–809. IEEE, 2020.

[9] Midas Nouwens, Ilaria Liccardi, Michael Veale, David Karger, and Lalana Kagal. Dark Patterns after the GDPR: Scraping Consent Pop-Ups and Demonstrating Their Influence. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, CHI '20. ACM, 2020.

[10] Dominique Machuletz and Rainer Böhme. Multiple Purposes, Multiple Problems: A User Study of Consent Dialogs after GDPR. *Proceedings on Privacy Enhancing Technologies*, (2):481–498, 2020.

[11] Hana Habib, Sarah Pearman, Jiamin Wang, Yixin Zou, Alessandro Acquisti, Lorrie Faith Cranor, Norman Sadeh, and Florian Schaub. "It's

a Scavenger Hunt": Usability of Websites' Opt-Out and Data Deletion Choices. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, CHI '20. ACM, 2020.

[12] Sean O'Connor, Ryan Nurwono, and Eleanor Birrell. (Un)clear and (In)conspicuous: The right to opt-out of sale under CCPA, 2020.

[13] Maximilian Hils, Daniel W Woods, and Rainer Böhme. Measuring the Emergence of Consent Management on the Web. In *Proceedings of the Internet Measurement Conference 2020*, IMC '20. ACM, 2020.

[14] Daniel W Woods and Rainer Böhme. The commodification of consent. In *20th Annual Workshop on the Economics of Information Security, WEIS*, 2020.

[15] Gunes Acar, Christian Eubank, Steven Englehardt, Marc Juarez, Arvind Narayanan, and Claudia Diaz. The Web Never Forgets: Persistent Tracking Mechanisms in the Wild. In *Proceedings of the 2014 ACM SIGSAC Conference on Computer and Communications Security*, CCS '14, pages 674–689. ACM, 2014.

[16] Steven Englehardt, Dillon Reisman, Christian Eubank, Peter Zimmerman, Jonathan Mayer, Arvind Narayanan, and Edward W. Felten. Cookies That Give You Away: The Surveillance Implications of Web Tracking. In *Proceedings of the 24th International Conference on World Wide Web*, WWW '15, pages 289–299, Republic and Canton of Geneva, CHE, 2015. International World Wide Web Conferences Steering Committee.

[17] Steven Englehardt and Arvind Narayanan. Online Tracking: A 1-Million-Site Measurement and Analysis. In *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security*, CCS '16, pages 1388–1401. ACM, 2016.

[18] Pierre Laperdrix, Nataliia Bielova, Benoit Baudry, and Gildas Avoine. Browser Fingerprinting: A Survey. *ACM Trans. Web*, 14(2), April 2020.

[19] T. Bujlow, V. Carela-Español, J. Solé-Pareta, and P. Barlet-Ros. A Survey on Web Tracking: Mechanisms, Implications, and Defenses. *Proceedings of the IEEE*, 105(8):1476–1510, 2017.

[20] Simon Byers, Lorrie Faith Cranor, and David Kormann. Automated analysis of P3P-enabled web sites. In *Proceedings of the 5th International Conference on Electronic Commerce*, pages 326–338, 2003.

[21] Patricia Beatty, Ian Reay, Scott Dick, and James Miller. P3P adoption on e-commerce web sites: a survey and analysis. *IEEE Internet Computing*, 11(2):65–71, 2007.

[22] Ian Reay, Patricia Beatty, Scott Dick, and James Miller. Privacy policies and national culture on the internet. *Information Systems Frontiers*, 15(2):279–292, 2013.

[23] Riva Richmond. A loophole big enough for a cookie to fit through. *New York Times*, 2010. https://nyti.ms/2mDvTBQ.

[24] Lorrie Faith Cranor, Manjula Arjula, and Praveen Guduru. Use of a P3P user agent by early adopters. In *Proceedings of the 2002 ACM Workshop on Privacy in the Electronic Society*, pages 1–10, 2002.

[25] World Wide Web Consortium. Tracking Protection Working Group. https://www.w3.org/2011/tracking-protection/, 2011.

[26] Julia Angwin. Microsoft's "Do Not Track" Move Angers Advertising Industry. https://www.wsj.com/articles/BL-DGB-24506, 2012.

[27] Chrome Blog. Longer battery life and easier website permissions. https://chrome.googleblog.com/2012/11/longer-battery-life-and-easier-website.html, 2012.

[28] Future of Privacy Forum. Companies that have implemented Do Not Track. https://allaboutdnt.com/companies/, 2020.

[29] Alex Fowler. Mozilla's new Do Not Track dashboard: Firefox users continue to seek out and enable DNT. https://blog.mozilla.org/netpolicy/2013/05/03/mozillas-new-do-not-track-dashboard-firefox-users-continue-to-seek-out-and-enable-dnt/, 2013.

[30] Robin Berjon, Sebastian Zimmeck, Ashkan Soltani, David Harbage, and Peter Synder. Global Privacy Control (GPC) Unofficial Draft 15 October 2020. https://globalprivacycontrol.github.io/gpc-spec/, 2020.

[31] IAB Europe. What is the Transparency and Consent Framework (TCF)? https://iabeurope.eu/transparency-consent-framework/, 2020.

[32] J. R. Mayer and J. C. Mitchell. Third-party web tracking: Policy and technology. In *2012 IEEE Symposium on Security and Privacy*, pages 413–427. IEEE, 2012.

[33] Balachander Krishnamurthy and Craig E Wills. On the leakage of personally identifiable information via online social networks. In *Proceedings of the 2nd ACM workshop on online social networks*, pages 7–12, 2009.

[34] Gunes Acar, Steven Englehardt, and Arvind Narayanan. No boundaries: data exfiltration by third parties embedded on web pages. *Proceedings on Privacy Enhancing Technologies*, 2020(4):220 – 238, 2020.

[35] Shehroze Farooqi, Maaz Musa, Zubair Shafiq, and Fareed Zaffar. Canary-trap: Detecting data misuse by third-party apps on online social networks. *Proceedings on Privacy Enhancing Technologies*, 2020(4):336 – 354, 2020.

[36] Irwin Reyes, Primal Wijesekera, Joel Reardon, Amit Elazari Bar On, Abbas Razaghpanah, Narseo Vallina-Rodriguez, and Serge Egelman. "Won't Somebody Think of the Children?" Examining COPPA Compliance at

Scale. *Proceedings on Privacy Enhancing Technologies*, 2018(3):63 – 83, 2018.

[37] Hamza Saleem and Muhammad Naveed. SoK: Anatomy of Data Breaches. *Proceedings on Privacy Enhancing Technologies*, 2020(4):153 – 174, 2020.

[38] Sébastien Henri, Gines Garcia-Aviles, Pablo Serrano, Albert Banchs, and Patrick Thiran. Protecting against Website Fingerprinting with Multihoming. *Proceedings on Privacy Enhancing Technologies*, 2020(2):89 – 110, 01 Apr. 2020.

[39] Miti Mazmudar and Ian Goldberg. Mitigator: Privacy policy compliance using trusted hardware. *Proceedings on Privacy Enhancing Technologies*, 2020(3):204 – 221, 2020.

[40] Martino Trevisan, Stefano Traverso, Eleonora Bassi, and Marco Mellia. 4 Years of EU Cookie Law: Results and Lessons Learned. *Proceedings on Privacy Enhancing Technologies*, 2019(2):126 – 145, 2019.

[41] Daniel W. Woods and Rainer Böhme. SoK: Quantifying cyber risk. In *IEEE Symposium on Security and Privacy*, May 2021.

[42] Laura Shipp and Jorge Blasco. How private is your period?: A systematic analysis of menstrual app privacy policies. *Proceedings on Privacy Enhancing Technologies*, 2020(4):491 – 510, 2020.

[43] Ryan Amos, Gunes Acar, Elena Lucherini, Mihir Kshirsagar, Arvind Narayanan, and Jonathan Mayer. Privacy Policies over Time: Curation and Analysis of a Million-Document Dataset. *arXiv preprint arXiv:2008.09159*, 2020.

[44] Martin Degeling, Christine Utz, Christopher Lentzsch, Henry Hosseini, Florian Schaub, and Thorsten Holz. We Value Your Privacy ... Now Take Some Cookies: Measuring the GDPR's Impact on Web Privacy. In *26th Annual Network and Distributed System Security Symposium*, NDSS '19. The Internet Society, 2019.

[45] Thomas Linden, Rishabh Khandelwal, Hamza Harkous, and Kassem Fawaz. The Privacy Policy Landscape After the GDPR. *Proceedings on Privacy Enhancing Technologies*, 2020(1):47 – 64, 01 Jan. 2020.

[46] Judith S Olson, Jonathan Grudin, and Eric Horvitz. A study of preferences for sharing and privacy. In *CHI'05 extended abstracts on Human factors in Computing Systems*, pages 1985–1988, 2005.

[47] Mark S Ackerman, Lorrie Faith Cranor, and Joseph Reagle. Privacy in e-commerce: examining user scenarios and privacy preferences. In *Proceedings of the 1st ACM Conference on Electronic commerce*, pages 1–8, 1999.

[48] Ben Weinshel, Miranda Wei, Mainack Mondal, Euirim Choi, Shawn Shan, Claire Dolin, Michelle L. Mazurek, and Blase Ur. Oh, the Places You've Been! User Reactions to Longitudinal Transparency About Third-Party Web Tracking and Inferencing. In *Proceedings of the 2019 ACM SIGSAC Conference on Computer and Communications Security*, CCS '19, pages 149–166. ACM, 2019.

[49] Sarah Spiekermann, Jens Grossklags, and Bettina Berendt. E-Privacy in 2nd Generation E-Commerce: Privacy Preferences versus Actual Behavior. In *Proceedings of the 3rd ACM Conference on Electronic Commerce*, EC '01, pages 38–47. ACM, 2001.

[50] Susanne Barth and Menno DT De Jong. The privacy paradox – Investigating discrepancies between expressed privacy concerns and actual online behavior – A systematic literature review. *Telematics and informatics*, 34(7):1038–1058, 2017.

[51] Nina Gerber, Paul Gerber, and Melanie Volkamer. Explaining the privacy paradox: A systematic review of literature investigating privacy attitude and behavior. *Computers & Security*, 77:226–261, 2018.

[52] Lorrie Faith Cranor. P3P: Making privacy policies more useful. *IEEE Security & Privacy*, 1(6):50–55, 2003.

[53] Rakesh Agrawal, Jerry Kiernan, Ramakrishnan Srikant, and Yirong Xu. XPref: a preference language for P3P. *Computer Networks*, 48(5):809 – 827, 2005. Web Security.

[54] Johnson Iyilade and Julita Vassileva. P2U: a privacy policy specification language for secondary data sharing and usage. In *2014 IEEE Security and Privacy Workshops*, pages 18–22. IEEE, 2014.

[55] Jean Yang, Kuat Yessenov, and Armando Solar-Lezama. A language for automatically enforcing privacy policies. *ACM SIGPLAN Notices*, 47(1):85–96, 2012.

[56] Monir Azraoui, Kaoutar Elkhiyaoui, Melek Önen, Karin Bernsmed, Anderson Santana De Oliveira, and Jakub Sendor. A-PPL: An Accountability Policy Language. In *Data Privacy Management, Autonomous Spontaneous Security, and Security Assurance*, pages 319–326, Cham, 2015. Springer.

[57] Lalana Kagal, Chris Hanson, and Daniel Weitzner. Using dependency tracking to provide explanations for policy management. In *2008 IEEE Workshop on Policies for Distributed Systems and Networks*, pages 54–61. IEEE, 2008.

[58] Ponnurangam Kumaraguru, Lorrie Cranor, Jorge Lobo, and Seraphin Calo. A survey of privacy policy languages. In *Workshop on Usable IT Security Management (USM 07): Proceedings of the 3rd Symposium on Usable Privacy and Security, ACM*, 2007.

[59] Jun Zhao, Reuben Binns, Max Van Kleek, and Nigel Shadbolt. Privacy languages: Are we there yet to enable user controls? In *Proceedings of the 25th International Conference Companion on World Wide Web*, WWW '16 Companion, pages 799–806. International World Wide Web Conferences Steering Committee, 2016.

[60] Saffija Kasem-Madani and Michael Meier. Security and privacy policy languages: A survey, categorization and gap identification. *CoRR*, abs/1512.00201, 2015.

[61] Victor Morel and Raúl Pardo. SoK: Three facets of privacy policies. In *WPES'20: Proceedings of the 19th Workshop on Privacy in the Electronic Society, Virtual Event, USA, November 9, 2020*, pages 41–56. ACM, 2020.

[62] Lorrie Faith Cranor, Serge Egelman, Steve Sheng, Aleecia M McDonald, and Abdur Chowdhury. P3P deployment on websites. *Electronic Commerce Research and Applications*, 7(3):274–293, 2008.

[63] Ian Reay, Scott Dick, and James Miller. An analysis of privacy signals on the World Wide Web: Past, present and future. *Inf. Sci.*, 179(8):1102–1115, 2009.

[64] Célestin Matte, Cristiana Santos, and Nataliia Bielova. Purposes in IAB Europe's TCF: which legal basis and how are they used by advertisers? In *Annual Privacy Forum*, 2020.

[65] Yee-Lin Lai and Kai-Lung Hui. Internet opt-in and opt-out: Investigating the roles of frames, defaults and privacy concerns. In *Proceedings of the 2006 ACM SIGMIS CPR Conference on Computer Personnel Research*, SIGMIS CPR '06, pages 253–263. ACM, 2006.

[66] Rainer Böhme and Stefan Köpsell. Trained to accept? A field experiment on consent dialogs. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '10, pages 2403–2406. ACM, 2010.

[67] Idris Adjerid, Alessandro Acquisti, Laura Brandimarte, and George Loewenstein. Sleights of privacy: Framing, disclosures, and the limits of transparency. In *Proceedings of the Ninth Symposium on Usable Privacy and Security*, SOUPS '13. ACM, 2013.

[68] Barry M Leiner, Vinton G Cerf, David D Clark, Robert E Kahn, Leonard Kleinrock, Daniel C Lynch, Jon Postel, Larry G Roberts, and Stephen Wolff. A brief history of the internet. *ACM SIGCOMM Computer Communication Review*, 39(5):22–31, 2009.

[69] Mehdi Nikkhah, Aman Mangal, Constantine Dovrolis, and Roch Guérin. A statistical exploration of protocol adoption. *IEEE/ACM Transactions on Networking*, 25(5):2858–2871, 2017.

[70] Jakub Czyz, Mark Allman, Jing Zhang, Scott Iekel-Johnson, Eric Osterweil, and Michael Bailey. Measuring IPv6 adoption. *SIGCOMM Comput. Commun. Rev.*, 44(4):87–98, August 2014.

[71] Xuequn Wang and Sebastian Zander. Extending the model of internet standards adoption: A cross-country comparison of IPv6 adoption. *Information & Management*, 55(4):450 – 460, 2018.

[72] M. Nikkhah and R. Guérin. Migrating the Internet to IPv6: An Exploration of the When and Why. *IEEE/ACM Transactions on Networking*, 24(4):2291–2304, 2016.

[73] Ralph Holz, Lothar Braun, Nils Kammenhuber, and Georg Carle. The SSL Landscape: A Thorough Analysis of the x.509 PKI Using Active and Passive Measurements. In *Proceedings of the 2011 ACM SIGCOMM Conference on Internet Measurement Conference*, IMC '11, pages 427–444. ACM, 2011.

[74] Andy Ozment and Stuart E Schechter. Bootstrapping the adoption of internet security protocols. In *5th Annual Workshop on the Economics of Information Security, WEIS*, 2006.

[75] Adrienne Porter Felt, Richard Barnes, April King, Chris Palmer, Chris Bentzel, and Parisa Tabriz. Measuring HTTPS adoption on the web. In *Proceedings of the USENIX Security Symposium (USENIX Security 17)*, pages 1323–1338, 2017.

[76] Victor Le Pochat, Tom van Goethem, Samaneh Tajalizadehkhoob, Maciej Korczynski, and Wouter Joosen. Tranco: A research-oriented top sites ranking hardened against manipulation. In *26th Annual Network and Distributed System Security Symposium*, NDSS '19. The Internet Society, 2019.

[77] Symantec. Symantec RuleSpace: URL categorization database, 2020.

[78] Iskander Sanchez-Rola, Matteo Dell'Amico, Platon Kotzias, Davide Balzarotti, Leyla Bilge, Pierre-Antoine Vervier, and Igor Santos. Can I Opt Out Yet? GDPR and the Global Illusion of Cookie Control. In *Proceedings of the 2019 ACM Asia Conference on Computer and Communications Security*, Asia CCS '19, pages 340–351. ACM, 2019.

[79] Pelayo Vallina, Victor Le Pochat, Álvaro Feal, Marius Paraschiv, Julien Gamba, Tim Burke, Oliver Hohlfeld, Juan Tapiador, and Narseo Vallina-Rodriguez. Mis-shapes, Mistakes, Misfits: An Analysis of Domain Classification Services. In *Proceedings of the Internet Measurement Conference 2020*, IMC '20. ACM, 2020.

[80] Mozilla Foundation. Public suffix list. https://publicsuffix.org/, 2007–2020.

[81] Yana Dimova, Gunes Acar, Lukasz Olejnik, Wouter Joosen, and Tom van Goethem. The CNAME of the Game: Large-scale Analysis of DNS-based Tracking Evasion. *Proceedings on Privacy Enhancing Technologies*, 2021.

[82] Inside Privacy. Digital Advertising Alliance Leaves Do Not Track Group. https://www.insideprivacy.com/advertising-marketing/digital-advertising-alliance-leaves-do-not-track-group-2/, 2013.

[83] IAB Tech Lab. Global Privacy Working Group. https://iabtechlab.com/working-groups/global-privacy-working-group/, 2011.

[84] Andrew L Russell. 'Rough consensus and running code' and the Internet-OSI standards war. *IEEE Annals of the History of Computing*, 28(3):48–61, 2006.

[85] Christopher Soghoian. The History of the Do Not Track Header. http://paranoia.dubfire.net/2011/01/history-of-do-not-track-header.html, 2011.

[86] Carl Shapiro, Shapiro Carl, Hal R Varian, et al. *Information rules: a strategic guide to the network economy*. Harvard Business Press, 1998.

[87] Kochava Inc. Quantcast and Kochava Partnership Delivers Combined Web and Mobile App Solution for CCPA. https://www.businesswire.com/news/home/20200207005054/en/Quantcast-and-Kochava-Partnership-Delivers-Combined-Web-and-Mobile-App-Solution-for-CCPA, 2018.

[88] Johnny Ryan. Regulatory complaint concerning massive, web-wide data breach by Google and other "ad tech" companies under Europe's GDPR. https://brave.com/adtech-data-breach-complaint/, 2018.

[89] Natasha Lomas. Brave Accueses European governments of GDPR re-sourcing failure. https://techcrunch.com/2020/04/27/brave-accuses-european-governments-of-gdpr-resourcing-failure/, 2020.

[90] Johnny Ryan. Formal GDPR complaint against IAB Europe's "cookie wall" and GDPR consent guidance. https://brave.com/iab-cookie-wall/, 2019.

[91] Tue Goldschmieding. New important decision on cookies from the Danish Data Protection Agency. https://gorrissenfederspiel.com/en/knowledge/news/new-important-decision-on-cookies-from-the-danish-data-protection-agency, 2020.

[92] Aaron Ceross and Andrew Simpson. Rethinking the Proposition of Privacy Engineering. In *Proceedings of the New Security Paradigms Workshop*, NSPW '18, pages 89–102. ACM, 2018.

[93] Carl Shapiro and Hal R Varian. The art of standards wars. *California Management Review*, 41(2):8–32, 1999.

[94] Christoph Bösch, Benjamin Erb, Frank Kargl, Henning Kopp, and Stefan Pfattheicher. Tales from the dark side: Privacy dark strategies and privacy dark patterns. *Proceedings on Privacy Enhancing Technologies*, 2016(4):237–254, 2016.

[95] Arunesh Mathur, Gunes Acar, Michael J Friedman, Elena Lucherini, Jonathan Mayer, Marshini Chetty, and Arvind Narayanan. Dark patterns at scale: Findings from a crawl of 11k shopping websites. *Proceedings of the ACM on Human-Computer Interaction*, 3(CSCW):1–32, 2019.

[96] Arvind Narayanan, Arunesh Mathur, Marshini Chetty, and Mihir Kshirsagar. Dark Patterns: Past, Present, and Future. *ACM Queue*, 18(2):67–92, 2020.

APPENDIX

Table 4.3: Regression Coefficients for TCF 2.x Adoption

	Dependent variable:		
	TCF 2.x Adoption		
	(1)	(2)	(3)
Google Ads	0.150^{***}	0.151^{***}	3.502^{***}
	(0.043)	(0.044)	(0.140)
log(# contacted SLDs)	0.765^{***}	0.596^{***}	1.787^{***}
	(0.020)	(0.020)	(0.053)
Category: Business		-0.436^{***}	-0.430^{***}
		(0.055)	(0.055)
Category: Education		-1.384^{***}	-1.385^{***}
		(0.098)	(0.098)
Category: Government		-2.479^{***}	-2.506^{***}
		(0.303)	(0.304)
Category: News & Entertainment		0.954^{***}	0.994^{***}
		(0.031)	(0.031)
Category: Shopping		-0.885^{***}	-0.826^{***}
		(0.067)	(0.067)
Category: Technology		-0.487^{***}	-0.469^{***}
		(0.055)	(0.055)
Google Ads * log(# contacted SLDs)			-1.517^{***}
			(0.058)
Constant	-4.614^{***}	-4.189^{***}	-6.558^{***}
	(0.045)	(0.048)	(0.121)
Observations	92,001	82,326	82,326
McFadden's Pseudo-R^2	0.08	0.13	0.14
Note:			*p<0.1; **p<0.05; ***p<0.01

Table 4.4: Summary Statistics

Variable	N	Min	Mean	Max
TCF 2.x Adoption	92,475	0	0.072	1
Google Ads	92,538	0	0.570	1
log(# contacted SLDs)	92,538	0	2.189	5.004
Category: Business	88,269	0	0.114	1
Category: Education	88,269	0	0.078	1
Category: Government	88,269	0	0.034	1
Category: News & Entertainment	88,269	0	0.210	1
Category: Shopping	88,269	0	0.086	1
Category: Technology	88,269	0	0.132	1

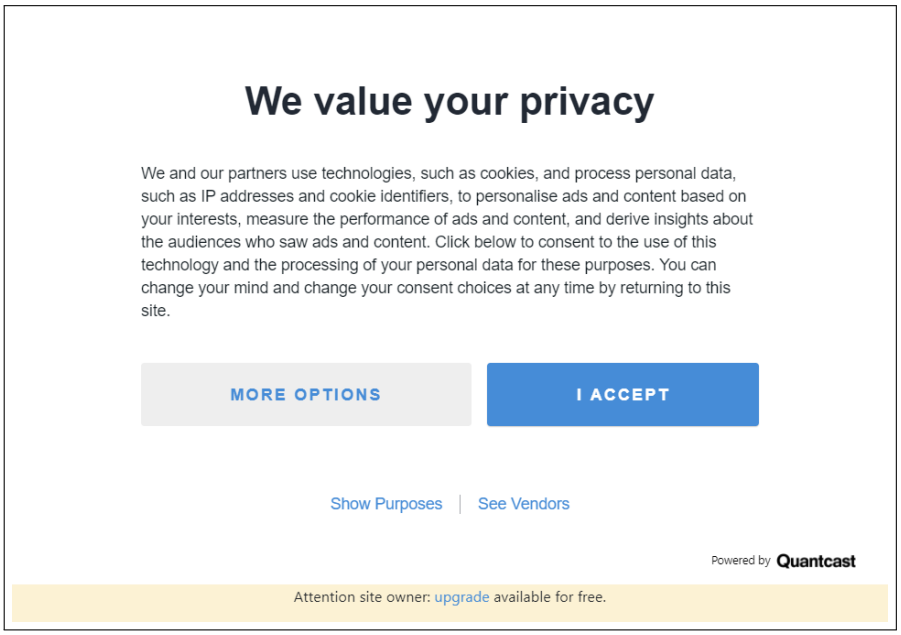


Figure 4.13: Starting August 5th 2020, Quantcast added a prominent deprecation message at the bottom of all its customers' TCF 1.x consent dialogs, prompting them to switch to TCF 2.0.

5

MEASURING THE EMERGENCE OF CONSENT MANAGEMENT ON THE WEB

AUTHORS

Maximilian Hils, University of Innsbruck
Daniel Woods, University of Innsbruck
Rainer Böhme, University of Innsbruck

CONFERENCE

ACM Internet Measurement Conference 2020 (IMC)
27–29 October 2020, Virtual Event.

ABSTRACT

Privacy laws like the General Data Protection Regulation (GDPR) and the California Consumer Privacy Act (CCPA) have pushed internet firms processing personal data to obtain user consent. Uncertainty around sanctions for non-compliance led many websites to embed a Consent Management Provider (CMP), which collects users' consent and shares it with third-party vendors and other websites. Our paper maps the formation of this ecosystem using longitudinal measurements. Primary and secondary data sources are used to measure each actor within the ecosystem. Using 161 million browser crawls, we estimate that CMP adoption doubled from June 2018 to June 2019 and then doubled again until June 2020. Sampling 4.2 million unique domains, we observe that CMP adoption is most prevalent among moderately popular websites (Tranco top 50-10k) but a long tail exists. Using APIs from the ad-tech industry, we quantify the purposes and lawful bases used to justify processing personal data. A controlled experiment on a public website provides novel insights into how the time-to-complete of two leading CMPs' consent dialogues varies with the preferences expressed, showing how privacy aware users incur a significant time cost.

5.1 INTRODUCTION

Vendors harvesting personal data prefer operating beyond the user's attention as evidenced by the use of secret tracking technologies [38, 1, 29]. This was tolerated by websites who rely on advertising revenues [51]. Sanctions associated with recent privacy laws threaten this state of affairs. In the EU, the General Data Protection Regulation (GDPR) requires firms processing personal data to establish a legal basis, such as by obtaining user consent. In the US, the California Consumer Privacy Act (CCPA) requires websites to collect the consent of minors and also to allow users to opt-out of the sale of their personal data. To comply with both laws, an infrastructure of consent must be designed so that users can consent to the privacy practices of websites and Ad-tech vendors.

In the past, each website offered a unique privacy policy and dialogue. This diversity overwhelmed users who could not commit hundreds of hours to reading each privacy policy [36, 6] nor navigate novel interface designs without making errors [2]. Privacy advocates argued that users should set preferences in the browser to avoid such problems [9, 27, 34], whereas Ad-tech companies lobbied against standardized privacy. However, the new imperative to obtain consent creates problems for Ad-tech vendors who must manage and document heterogeneous forms of consent collected across multiple websites.

Consent management providers (CMPs) emerged in the last three years to standardize the collection of online consent. These intermediaries define legal terms and conditions, present these to users via an embedded consent dialogue, store the resulting signal, and share it with third-parties. In essence, CMPs have created a consent ecosystem involving users, websites, and third-party vendors. For example, one CMP allows websites to collect consent for a 'Global Vendor List' with a membership fee of 1200 €, which was termed the commodification of consent [60].

The rise of CMPs represents a new stage in how privacy preferences are communicated, with previous stages including cookies settings in browsers [37] or custom cookie banners on websites [53]. This paper offers a longitudinal study of the formation of a consent ecosystem orchestrated by CMPs. We introduce the notion of a consent flow—from users through consent dialogues to a website and then onto third-parties—and make measurements at each interface. This complements post-GDPR related work relying on snapshots of relatively small samples of domains, which is shown in Figure 5.1.

Our insights include:

- Using 161 million browser crawls, we measure CMP adoption over time and by website popularity. We show that uptake is most prevalent among 'mid-market' sites ($50^{\text{th}} - 10,000^{\text{th}}$), although this varies between CMPs. We also show the winners and losers of inter-CMP competition in the form of websites switching CMPs.

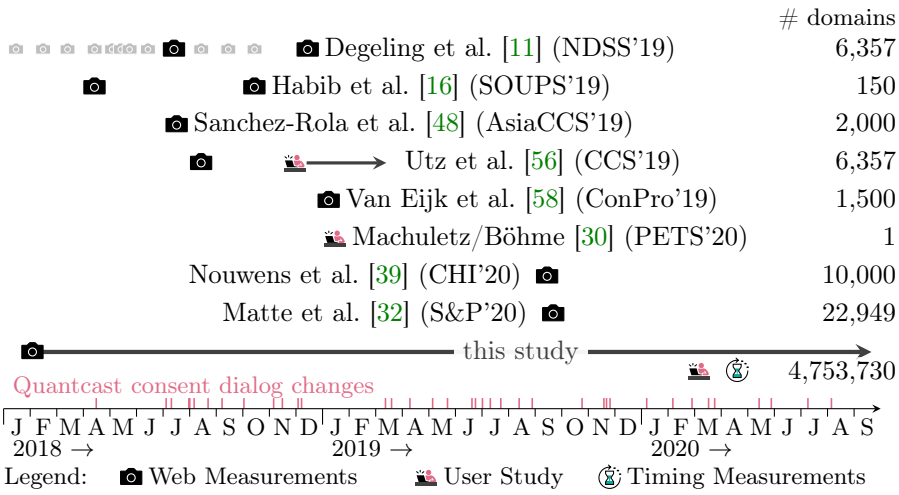


Figure 5.1: Previous studies conducted point-in-time snapshots of small samples in a rapidly changing environment. For example, the consent prompt of a single CMP (Quantcast) changed 38 times in our observation period.

- In terms of methodology, we introduce a novel URL sampling approach seeded by social media shares, which improves subsite coverage. This is complemented by a traditional toplist sample.
- Using APIs from the Ad-tech industry, we quantify the purposes and lawful bases used to justify processing personal data. We find many vendors claiming 'legitimate interest', which allows them to process data without the user's consent.
- We address gaps in the literature by measuring the time to complete consent dialogues, highlighting how users incur a significant time cost when opting out.

Section 5.2 provides information about the consent ecosystem. Section 5.3 describes our measurement approach. Section 5.4 presents our results, which are discussed in Section 5.5. We identify related work in Section 5.6 and offers conclusions in Section 5.7.

5.2 BACKGROUND

Section 5.2.1 describes how privacy laws create demand for consent management. Section 5.2.2 describes the organisations and technical standards relevant to consent management solutions.

5.2.1 *Privacy Laws and Consent*

The role of user consent in recent privacy laws is the most significant aspect for this paper. The GDPR applies to all firms processing personal data, which entangles Ad-tech trackers and data brokers as well as websites. Such firms can establish a legal basis for doing so by obtaining user consent (Article 6.1a) or

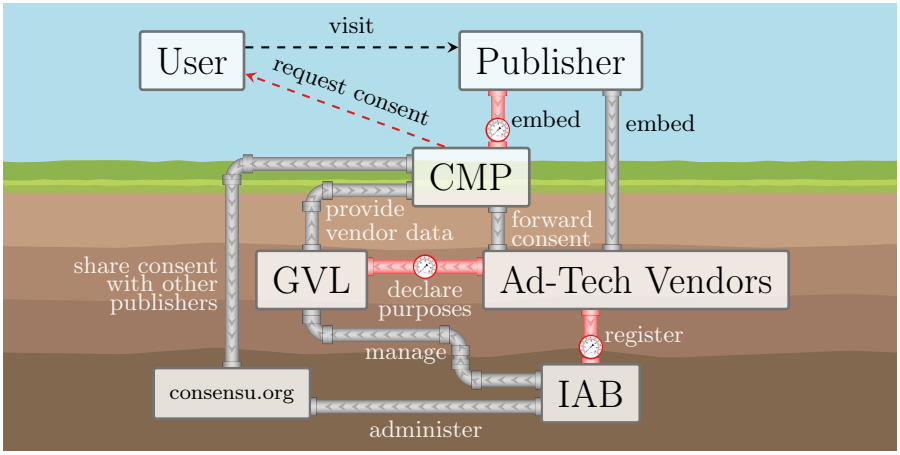


Figure 5.2: Surfacing the web's new compliance engine: Publishers embed CMPs, which display consent prompts to users, forward consent decisions to ad-tech vendors and also share it globally across websites. In the background, the IAB orchestrates this through its Transparency and Consent Framework (TCF).

by claiming a legitimate interest (Article 6.1b–f), such as if the data processing protects the "vital interests of the data subject or of another natural person" (6.1d) [47]. If controllers choose to obtain consent, it must be a "freely given, specific, informed and unambiguous indication of the data subject's wishes" (Recital 32) and "documented" (7.1). A data controller infringing either Article 6 or 7 is punishable by "a fine up to €20 million or up to 4% of the annual worldwide revenue."

In the United States, the California Consumer Privacy Act, which came into effect in January 2020, requires websites to: obtain parental consent for users under 13; affirmative consent for those under 16; and to allow other users to opt-out of the sale of their personal data [17]. The CCPA and GDPR further differ in the obligations on third-party vendors and the definition of personal information. The resulting uncertainty created a business opportunity for CMPs who claim to specialize in compliance. The next section describes the resulting products.

5.2.2 *Consent Management Solutions*

Ambiguity about how to technically implement the principles of privacy law [7] led to heterogeneity in consent management solutions. In response, the Internet Advertising Bureau (IAB) – not to be confused with the Internet Architecture Board – developed the Transparency and Consent Framework (TCF), "the only GDPR consent solution built by the industry for the industry" [20]. The TCF standardizes and centralizes the storage of 'global' consent cookies. It is visualized in Figure 5.2. We describe this technical standard to illustrate what CMPs do, and also because it is implemented by many but not all of the CMPs we measure in later sections.

The first building block of the TCF is the definition of purposes and features that are shown to users. In TCF 1.0, purposes define reasons for collecting personal data, for example; personalization, ad selection, or usage analytics. Features on the other hand describe methods of data use that overlap multiple purposes, such as combination with offline sources. A full list of purposes and features can be found in Table 5.2. Both must be disclosed to the users, but users are only given control over consenting to individual purposes.

The second building block of the TCF is the Global Vendor List (GVL), a master list of advertisers participating in the framework. The GVL is maintained by the IAB. Vendors declare the purposes for which they collect data and the features upon which they rely. They can also declare legitimate interest for specific purposes, which allows them to process personal data under the GDPR even if the user does not consent. For each advertiser, the GVL contains; a name, a link to the advertiser's privacy policy, the feature and purpose ids consent is requested for, and the declared legitimate interests. Registered advertisers pay a yearly management fee of 1.200 €. Cookie prompts implementing the TCF often request consent for all advertisers in this list, even though the website does not have a business relationship with every vendor. If the list is updated with new vendors (or additional purposes), users are prompted with a new dialogue in order to obtain additional consent.

The third building block involves the *Consent Management Providers* implementing the TCF on publishers' websites. They provide the cookie prompt, store the user's choice as a browser cookie, and provide an API for advertisers to access this information. The IAB also maintains a public list for CMPs, which lists 150 participating providers as of May 2020 [19]. A website wishing to implement the TCF independently must become a CMP, otherwise they can out-source this to an existing CMP. In reality, a handful of CMPs dominate the market.

Beyond the technical standard, IAB Europe also governs the surrounding ecosystem. The legal terms used in consent dialogues, such as the purposes of data collection, are standardized in the TCF. Firms adopting the standard are expected to follow the defined policy and IAB Europe publicizes a tool to audit CMPs (but not vendors). We go on to provide evidence that the TCF is inconsistently implemented in practice and not at all in some cases, such as for CMPs targeting the US market.

5.3 MEASUREMENT APPROACH

We identify our items of interest (what we want to measure) in Section 5.3.1. We map the items to a set of indicators and measurement methods that collectively describe our methodology in Section 5.3.2. Finally, we assess the threats to reliability and validity of our methodology in Section 5.3.5.

5.3.1 *Items of Interest*

The following items (**I1**–**I7**) span the consent ecosystem, which is visualized in Figure 5.2. In particular, the red arrows and pipes with pressure gauges are the links in the ecosystem that we measure. We know little about the prevalence of CMPs on the web. This complicates generalizing results about CMPs from snapshot samples of the most popular websites with size in the order of thousands, as was done in previous work [39, 32]. In order to build a fuller picture of the consent ecosystem, we ask: how does CMP adoption vary according to website popularity (**I1**), and related, how has this changed over time and been influenced by legal developments (**I2**).

The third item of interest relates to publisher behaviour: to what extent do websites customize the embedded CMP (**I3**). Privacy laws describe how consent can be legally collected, violations of which have been studied in [39, 32]. The responsibility for such violations is far from clear when a website embeds a CMP, which is especially true when the CMP allows the website to customize the embedded consent dialogue.

Turning to vendors processing personal data, there are many reasons why a vendor might do so. Obtaining consent is not the only lawful basis for data processing. The fourth and fifth items are; why are vendors collecting personal data (**I4**), and what is their legal basis for doing so (**I5**).

One aspect that has not been considered in existing research is the additional effort required to reject data processing compared to accepting it. In most experiments, artificial dialogues are pre-installed on the subject's machine or loaded from a single source. In practice, users may already be habituated to the standardized CMP dialogs, but dialogs may need to send consent decisions to multiple vendors which incurs additional waiting time. This motivates our items at the user-interface; how long does it take CMPs to distribute consent decisions (**I6**), and to what extent does the user's dialog interaction time vary depending on which privacy preferences are expressed (**I7**).

5.3.2 *Measurement Methodology*

LARGE-SCALE WEB MEASUREMENT To measure the prevalence of consent prompts longitudinally, we analyze automated browser crawls recorded by the Netograph web measurement platform[1] described in Figure 5.3. Netograph was not built exclusively for this research project and exhibits some unique properties compared to existing methods. Most prominently, instead of sampling from a particular toplist at one point in time, our crawlers are constantly seeded with new URLs shared on social media platforms.

This approach is not a design choice made specifically for our research, but useful in our context as measurements are not limited to a domain's landing page (`https://example.com/`) but also cover arbitrary subsites

[1] https://netograph.io/

Figure 5.3: The Netograph measurement platform collects a realtime stream of URLs shared on social media and crawls them using Google Chrome. Custom browser instrumentation extracts metadata such as HTTP requests and cookies. We match captures with CMP indicators and use the Tranco toplist to normalize website popularity.

(`https://example.com/foo?bar`). Recent work has shown that subsites show a significant different behavior and an increase of privacy-invasive techniques [55].

Netograph ingests a live feed of social media posts, extracts all URLs, and submits them into a capture queue. URLs are visited once within a couple of minutes after submission. Crawls are performed on virtual machines in US and EU data centers of a large public cloud provider. 50% of crawls are done from within the EU, each URL is assigned randomly. Websites are opened using Google Chrome on Linux with its current default user agent[2], a desktop resolution of 1024×800, and `en-US` as the preferred browser language. All other settings are set to their defaults: Third party cookies are allowed, the "Do not Track" HTTP header is not set, and Flash is disabled. Due to the large volume of URLs, Netograph crawls with relatively aggressive timeouts, which are discussed further in Section 5.3.5.

For every capture, Netograph collects the following data points using custom browser instrumentation. First, HTTP headers are logged for all requests and responses. Additionally, connection-related metadata such as IP addresses and TLS certificate chains are stored. For every domain in a capture, its relation to the main page, all cookies, IndexedDB, LocalStorage, SessionStorage and WebSQL records are saved. Finally, a screenshot of the visible area (without scrolling) is taken. Netograph does not store page contents due to storage constraints. All crawl data is stored in a central database, which can be queried using a custom API. As of May 2020, this database stores 177,868,171 captures or about 23 billion HTTP requests.

TOPLIST-BASED WEB MEASUREMENT To make comparisons with related work, we have set up an additional Netograph-based crawling infrastructure for this study based on an internet toplist. In our analysis, we use the top 10k entries from the Tranco list created on 30 January 2020[3], which aggregates the ranks from the lists provided by Alexa, Cisco Umbrella, Majestic, and Quantcast [44]. This sample size is in the order of magnitude of previous studies (see # domains in Figure 5.1).

We first converted the Tranco list of domains to a list of URLs that can be crawled. For each *domain*, we attempted to establish a TLS connection to www.*domain* on port 443 and validate the certificate hostname using Mozilla's trust store. If the certificate is valid, we used https://www.*domain*/ as the seed URL for crawls. Otherwise, we attempted to open a TCP connection on port 80 and used http://www.*domain*/ on success. If this also failed, we used http://*domain*/ as the seed URL. We repeated this process three times over a week in order to catch temporarily unavailable domains.

Next, we crawled every URL in the toplist six times in immediate succession: First, we visited the website from a European university network using our

[2] Currently `Mozilla/5.0 (X11; Linux x86_64) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/83.0.4103.61 Safari/537.36`.
[3] Available at `https://tranco-list.eu/list/K8JW`.

crawler's default configuration. Second, we repeated this capture with an extended timeout. Third and fourth, we also captured with both German and British English as the preferred browser language. Finally, we submitted the same URLs to Netograph's task queues in the US and EU cloud as a control group. We retried all unsuccessful captures three times over the span of a week.

For all toplist crawls, we additionally stored the browser's DOM tree including the computed CSS styles. We also recorded a full-page screenshot (including scrolling). These extended features are not stored for the social media dataset due to their storage requirements.

PREVALENCE AND CUSTOMIZATION OF CMPS (**I1**–**I3**) In the second part of our analysis, we measure the prevalence of CMPs using our crawl data. This involves extracting the final effective second-level domain (by which we want to count), detecting the CMP in use, and interpolating missing data. For this analysis we restrict ourselves to six CMPs: The five major players already identified by Nouwens et al. [39] and LiveRamp, a new entrant that launched in December 2019.

We measure the market share of CMPs by determining the number of domains they are active on. As about 11% of all crawls include top-level domain redirects, taking the domain from a seed URL would be imprecise. Instead, we extract the domain from the final website address as it would be shown in the browser's address bar. We normalize this domain to the effective second-level domain using the Public Suffix List [13], which contains all suffixes under which internet users can directly register names. For example, a capture may start with https://tinyurl.com/... as a seed URL, which redirects to https://foo.example.github.io/..., which we normalize to example.github.io.

To determine the CMP in use, we inspected the behavior of the six CMPs under study and created fingerprints for each CMP based on their HTTP request patterns, CSS selectors, and extracted text. For each CMP, we first recorded the network traffic of multiple websites where it was embedded and consulted the documentation provided by the CMP. Second, we assembled multiple fingerprints of varying specificity (for example, from concrete URLs to second-level domains) using manual analysis. To make sure that we did not miss any CMP dialogs, we searched for the GDPR phrases listed in [11] in our toplist crawls. We then checked the screenshots from our toplist crawls and discarded all fingerprints that yield false-positives. Finally, we verified that the remaining fingerprints work accurately for historic data using Netograph's captured screenshots. Using this approach, we were able to identify a unique hostname for each consent dialog framework as a robust indicator. For example, even though OneTrust deploys very different dialog designs with no shared JavaScript code or CSS classes, all of them perform HTTP requests to cdn.cookielaw.org on page load. We list our synthesized indicators in Table 5.3 for reproducibility.

Finally, we also need to take into account that the sampling frequency of a domain is not fixed in our main dataset as the crawler is seeded from social media posts only. Consequently, we may not see less popular domains for prolonged periods. We account for this in two ways. First, we interpolate missing observation periods if both boundary measurements are classified equally. For example, if we observed Quantcast on example.com a month ago and observe it again today, we assume that example.com kept using Quantcast as their CMP throughout this period. If the boundary measurements disagree, we do not assume the presence of the CMP in the intermediate period. Second, we account for the fact that our measurements are right-censored by fading out the presence of a CMP after 30 days if no new measurements have been made yet. For example, if a website was last measured a week before our analysis, we assume that they still use the same CMP; if the last measurement was made on February 1st, we assume no CMP presence as of March 1st. Finally, as we crawl with a fixed sampling frequency for our toplist-based measurements, we do not need to interpolate for this dataset.

AD-TECH VENDOR BEHAVIOR (**I4**–**I5**) Recall that Ad-tech vendors need to declare in the TCF for which data processing purposes they either request consent or claim legitimate interest. To assess the behavior of vendors, we systematically analyzed previous versions of the GVL and inspected them for longitudinal changes. In particular, we measure every instance when an Ad-tech vendor joins or leaves the GVL, claims a new purpose falls under legitimate interest, begins requesting consent for a new purpose, stops claiming either, or changes from collecting consent to claiming legitimate interest or the other way round.

TIME TO CONSENT (**I6**–**I7**) An aspect that has not been studied in the literature is the relative time taken to express different consent preferences. We aim to quantify this by embedding the dialogues offered by two leading CMPs, namely Quantcast and TrustArc. Using real dialogues in a field experiment improves ecological validity relative to studies using dialogues developed for research purposes in a lab experiment that result in a very different *feel* for the participants who are not browsing *normally*.

First, we measured how a seemingly small user interface change impacts the time it takes users to make a positive or negative consent decision. We embedded Quantcast's CMP dialog on a popular website on the public internet for a short period of time in two configurations: One with an explicit "Reject" button and one that included a "More Options" at the same position which would then lead to a reject button (see Figures 5.11 and 5.12). This design is motivated by the French data protection authority's guidelines, which demand a real choice between accepting or refusing cookies presented at the same level [10]. All other dialog settings were left to the default values: The consent prompt was shown as a modal dialog in the center of the screen, consent for all vendors on the GVL was requested, the "Accept" button was

colored more prominently, and the dialog was only shown to visitors from the EU. We then measured the page load time (`DOMContentLoaded`), the time the dialog appeared (`__cmp('ping',...)`[4]), and the time it was closed as well as the user's consent decision (`__cmp('getConsentData',...)`). We also checked for the existence of already existing global consent cookies by manually fetching `https://api.quantcast.mgr.consensu.org/CookieAccess`, which returns the users's existing Quantcast TCF cookie. Repeated visitors will not be counted as the CMP stores the first consent decision and no additional dialogs will be shown.

Second, we noticed that some CMP dialogs require extended processing time if users decide to opt out. For example, TrustArc consent prompts disappear immediately if one accepts cookies, but otherwise make the user wait for prolonged periods while opt-out requests are being sent to a hodgepodge of third parties. In our testing, opting out required users to wait tens of seconds, which could be skipped at any time by giving consent. To make sure that these observations were not a fluke, we repeatedly visited a website embedding the TrustArc dialog, automated the opt-out process with a custom Google Chrome extension, and collected all HTTP requests and timings.

5.3.3 *Research Ethics*

Our time-to-consent measurements were conducted on a website with real users, which raises ethical concerns as we did not ask for consent prior to measuring their interactions with consent notices. We did so to ensure non-biased results, which is supported by previous research on consent dialogs [56]. We ensured that we did not harm website visitors and their privacy. We address privacy issues by data minimization, i.e. we only collected a user's consent decision and the timings described in Section 5.3.2. The timings for a single page visit are linked using a random non-persistent id generated on page load. We do not create or store any persistent identifiers. While we believe that the second dialog design may not fulfill the requirements of the GDPR, the website we ran our experiments on did not perform any personal data collection irrespective of the user's consent decision.

5.3.4 *Data Sources*

Recall that Netograph's web crawlers are seeded with URLs posted on social media. More specifically, we ingest all URLs shared on Reddit and 1% of public Tweets using Twitter's sample feed[5]. Note that this does not mean we see 1% of URLs: each popular URL has multiple chances to be spotted in the sample feed as it is re-shared and retweeted. So in effect our URL sample skews heavily towards popular URLs. Overall, Twitter accounts for 80% of

[4] The `__cmp()` function is standardized as part of the IAB's Transparency & Consent Framework, see Matte et al. [32].

[5] `https://developer.twitter.com/en/docs/labs/sampled-stream/overview`

all URLs. We skip a URL if we have captured the same domain in the last hour or the precise URL in the last 48 hours. This applies to about 40% of all submitted URLs. Our records span March 2018–September 2020, starting before the inception of GDPR and also covering the introduction of CCPA.

To track the development of the global vendor list, we systematically downloaded all 215 previously published versions of the GVL from https://vendorlist.consensu.org/vXXX/vendor-list.json and verified their accuracy using the Internet Wayback Machine. Likewise, we collected the change history of Quantcast's consent dialog in the same way.

To measure how long it takes for users to make a consent decision, we embedded Quantcast's CMP dialog and our collection script on `mitmproxy.org` for a short period of time in May 2020. We logged about 120,000 timestamps. Importantly for generalizing, the website we hosted our experiment on caters to a very technical and privacy-concious audience.

For our second timing experiment, we measured the raw waiting time (not including user interaction) it takes to reject all tracking on `forbes.com`'s TrustArc consent dialog. Measurements were performed hourly for two weeks in May 2020. These measurements were made from a European university as the vantage point.

The relationship between our items of interest, data sources, and vantage points is summarized again in the Appendix (Table 5.5).

5.3.5 *Reliability and Validity*

SOCIAL MEDIA SAMPLE BIAS While existing research is mostly based on the Alexa and Tranco toplists, our measurement platform is seeded using URLs obtained from social media posts. An obvious issue with this setup is that URLs shared on social media are not a representative sample of the internet. One would reasonably expect YouTube videos to be shared more than `mastercard.com`. Hence our sample exhibits a different coverage error than typical toplist-based studies, which are not representative of the internet either. Additionally, our choice of social media data feeds is heavily skewed towards Western culture. We rectify this bias in part by grouping captures by their effective second-level domain. In other words, popular domains have a higher sampling frequency in our dataset, but equal weight.

MISSING DATA Another threat to validity is that some domains in the toplist have never been shared on social media. This affects 1021 domains in the Tranco 10k list. Of these 1021 domains, 305 were not reachable via HTTP or HTTPS at all in our toplist measurements, 4 did not return a valid HTTP response and 67 returned an HTTP error status code. 184 domains redirected to another domain and were counted as the redirect target. The overwhelming majority ($> 90\%$) of the remaining 461 domains can be considered internet infrastructure that is not directly accessed by users, such as CDNs.

Location	US ☁	EU ☁	EU University			
User Agent	🇺🇸	🇺🇸	🇺🇸	🇺🇸	🇬🇧	🇩🇪
Timing				⏱	⏱	⏱
OneTrust	341	368	403	412	412	414
Quantcast	173	207	225	229	230	233
TrustArc	107	118	152	157	154	156
Cookiebot	92	97	96	98	99	99
LiveRamp	8	9	14	14	14	14
Crownpeak	8	8	8	9	9	9
$\sum$	729	807	898	919	918	925
Coverage	79%	87%	97%	99%	99%	100%

Table 5.1: Occurence of CMPs on websites in the Tranco 10k measured from different vantage points.

SUBSITES In contrast to previous research, we crawl not only a domain's landing page but also arbitrary subsites given by the seed URLs. This increases the reliability of our results as it allows us to detect CMPs that are only present on specific subdomains or subsites. However, we also encounter individual pages that do not include a CMP. For example, some websites do not embed any external scripts on their privacy policy page. As a simple heuristic, we classify a website as using a CMP if the CMP is included in at least every third capture. For 99.8% of all domains, the daily share of CMP captures is either consistently below 5% or above 95%.

The remaining 0.2% of websites include a set of larger websites who change their behavior depending on the user's location, for example by complying with CCPA in the US but responding with `HTTP 451 Unavailable For Legal Reasons` to European visitors.

CRAWLER LOCATION Netograph crawls all URLs from virtual machines rented from a large public cloud provider. Half of all captures are done from the EU and the US respectively. This matches the recommendations made by Van Eijk et al. [58] to perform crawls from both inside and outside the EU for cookie consent notices. As shown in Table 5.1, we observe significantly more CMP adoption when crawling from the EU. This observation matches Van Eijk et al.'s finding on vantage point difference and can be explained by websites that only embed a CMP for EU visitors. Still, many websites choose to always embed their CMP framework but configure it to only show consent dialogs to EU visitors.

However, we found that not only the originating country, but also the type of address space has a significant influence on measurement results. As shown in Table 5.1, the use of public cloud infrastructure makes us miss about 10% of all CMP dialogs in the Tranco 10k. We manually inspected the sites in question and found that this is predominantly caused by anti-bot interstitial pages offered by popular CDNs. In contrast to the vantage point, the choice

of browser language settings did not have a significant effect on our web measurements.

Lastly, we re-iterate our overall point that longitudinal measurements matter for web privacy measurements: Looking at the same measurements in January 2020 (see Table 5.4), we see that only 70% of CMP usage is visible in our measurements from the US. The rise in coverage can be explained by the increasing adoption of CCPA in recent months.

CRAWLER TIMEOUTS Due to the large volume of URLs, Netograph runs crawls with relatively aggressive timeouts. To determine if a page has finished loading, it looks at frame load events from Chrome, the timing of requests, an idle timeout of five seconds and a total page timeout of 45 seconds. We note that crawls are done with heavy CPU utilization and a comparison with captures from the desktop might not be apt. In any case, our approach differs from smaller toplist-based measurements, which can afford much more relaxed timeouts. We quantify this change in Table 5.1: The timeouts employed by our measurement platform make us miss about 2% of CMP usage.

CHOICE OF TOPLIST To determine website popularity, we used the Tranco toplist [44]. Tranco aggregates results from other lists such as the Alexa toplist, is hardened against manipulation, less susceptible to daily fluctuations, and emphasizes reproducibility by providing permanent citable references. This decision is on line with recent related work on cookie consent [32]. While Urban et al. adapt the suggestion in the Tranco paper to remove all websites with the same TLD+1 [55], we do not perform this in our case as services may vary in their behavior across TLDs. For example, `amazon.com` shows a different consent prompt than the EU version of `amazon.co.uk` as of May 2020. A much more important factor which previous work has not elaborated on is the choice of toplist size. We show in the next section that different toplist sizes yield significantly different results.

CMP DETECTION We found our detection of CMPs to be robust despite heterogeneous CMP implementations on different websites. By looking at network traffic patterns we do not rely on any HTML or DOM parsing, which we found to be much more unreliable for analyses which we ultimately decided not to include in this paper. In particular, network patterns often allow us to detect the presence of CMPs even if the website's CMP configuration does not trigger a dialog, for example because we visit a EU-centric website from the US or vice-versa. However, we acknowledge that our detection accuracy and robustness is difficult to quantify. We have manually evaluated patterns on other candidate domains, patterns on specific HTTP requests, patterns on CSS selectors, and patterns on extracted text to make sure that we do not miss any CMP implementations. Additionally, we have used the Internet Wayback Machine to validate that our patterns match correctly on historic data. The only exception to this is a two-day period in July 2018 when

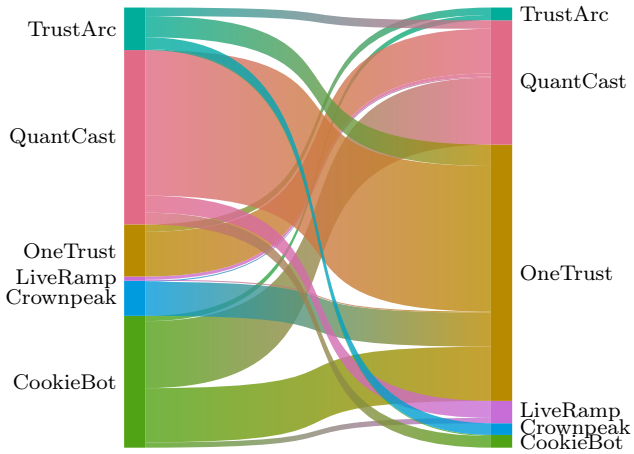


Figure 5.4: Sankey diagram of 414 CMP switches in the observation period (April 2018 to September 2020)

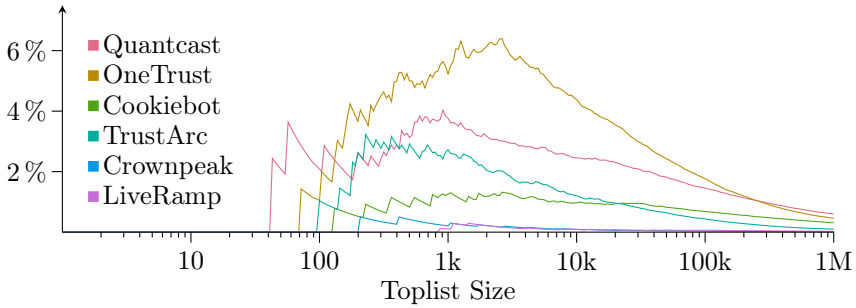


Figure 5.5: Cumulative CMP marketshare as a function of the toplist size (May 2020).

Quantcast embedded parts of their CMP script for all customers of their analytics service, a different line of the firm's business. We manually exclude this outlier in our calculations. We overcount if a website includes more than one CMP, but this only affects 0.01% of all captures.

5.4 RESULTS

This section is structured according to which part of the ecosystem we are focusing on; websites and CMPs in Section 5.4.1, vendors in Section 5.4.2, and the user-interface in Section 5.4.3.

5.4.1 *Measuring CMP Adoption*

Figure 5.5 shows how CMP adoption varies across the Tranco top million sites. The y-axis shows the percentage of firms embedding each CMP provider in the toplist with size corresponding to the x-axis. None of the largest websites embed the CMPs under consideration, likely because they have the in-house expertise to implement their own consent management solution. Speaking

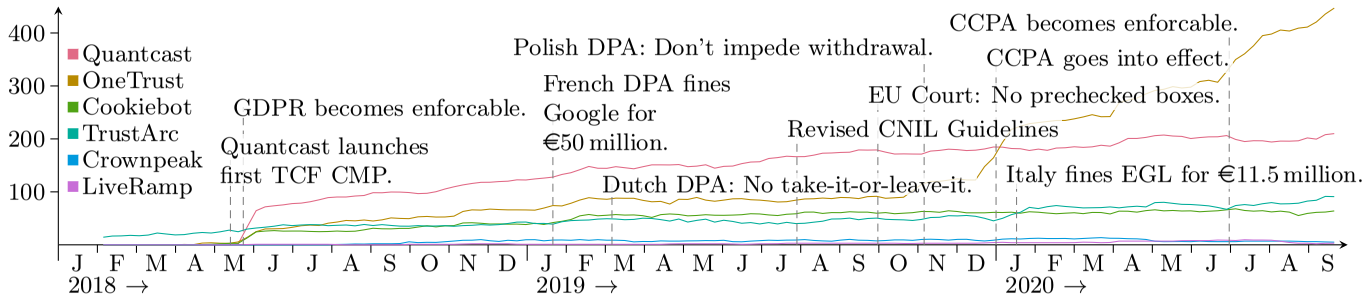


Figure 5.6: Number of websites in the Tranco 10k toplist that embed a CMP.
We include a non-exhaustive timeline of events with relevance to the GDPR and the CCPA.

to (**I1**), CMP adoption is most prevalent among the $50 - 10,000^{\text{th}}$ websites, especially in the top $1,000 - 5,000^{\text{th}}$ sites. Adoption tails off but never vanishes.

Interestingly, we see that different firms penetrate different sections of the market. For example, more of the top 100 sites embed Quantcast than the other CMP providers combined. However, OneTrust has the most customers among the $500 - 50,000^{\text{th}}$ sites, although Quantcast are more commonly adopted in the long tail.

Figure 5.6 shows how this has varied over time (**I2**). Laws like GDPR and CCPA coming into effect were significant drivers in CMP adoption, which suggests consent management solutions are more about regulatory compliance than improving user experience. However, events relevant to privacy law like fines or regulatory guidance do not affect adoption. Quantcast's solution is targeted at GDPR and they achieved market dominance early on, but their market growth slowed and was unaffected by the CCPA coming into effect. In contrast, OneTrust became the market leader by offering a flexible solution that could be tailored to the requirements of the CCPA. This can be seen in the share of sites with a EU+UK TLD for each CMP (Quantcast at 38.3% and OneTrust with 16.3%).

Our longitudinal approach can detect when websites change CMPs. Figure 5.4 describes the resulting dynamics. Quantcast and OneTrust both win and lose websites to each other. However, the true loser of inter-CMP competition is Cookiebot who have lost an order of magnitude more websites than they gained. The appendix contains further longitudinal insights by showing the CMP marketshare in January 2019, January 2020, and September 2020 (respectively Figure 5.14, 5.15 and 5.16). These three figures show how OneTrust over-hauled the early market dominance established by Quantcast.

We now turn to how publishers customize consent solutions (**I3**). CMPs differ in how much customizability they extend to publishers, we classify this into *closed customization* in which the publisher may choose between finitely many options, and *open customization* in which the publisher can choose infinitely many, such as via free-text fields. In addition, *publisher customization* occurs when the website implements consent management related functionality beyond that offered by the CMP. We characterize the observed customization for the three largest CMPs to illustrate the ways in which this varies. All reported statistics are based on our measurements from an EU university vantage point (see Table 5.1) where we have the browser's DOM tree and full page screenshots available for inspection.

Our sample includes 414 websites embedding OneTrust displaying a range of consent dialogues. The majority (61%) offer a conventional cookie banner with a 1-click accept button and a second button or link leading to a page with more information and fine-grained controls. Only 2.4% of the sites display a cookie banner containing an opt-out button with text like "Do Not Sell", "Reject/Manage Cookies", or "Deny All", although 40% of such banners require further clicks to confirm the opt-out. A minority (5.5%) of websites include a 'script banner' (cookie banners in all but name) with one "Accept" button and

one “Reject/Manage Scripts” button. Rather than showing any banner, 7.5% of the websites in our sample included a link to cookie or privacy information in the website footer. The link text was some variant of “Do Not Sell”, “California Privacy Rights”, or “Privacy Policy” in 11, 15 and 4 websites respectively. Two of the latter showed cookie banners only when accessed from a US IP.

Quantcast’s dialogues are more standardized. Barriers contain two buttons, the first of which allows the user to provide consent to the publisher and partners in one click. Closed customizability is offered as a choice between the second-button rejecting all or it leading to a second page with more-fine grained options. Of the 233 websites embedding Quantcast in our sample, 55% offer a 1-click reject all. The text on each button is an interesting example of open-customization and we find that 87% use some variation of "I agree/consent/accept", including non-English language translations. The publishers who do not (13%) use free-form texts including "Whatever", "Sounds good", and "Accept and move on" that may not qualify as affirmative consent.

TrustArc dialogues display more closed-customization in terms of button structure but have much less open-customization in terms of button wording. Of the 156 websites embedding TrustArc: 7% have a dialogue with a first-page button that instantly opts out; 12% have a first-page opt-out that must establish a connection with multiple partners (we measure the time to do so in Section 5.4.3); 44% include a first-page button that implies the user has autonomy; 31% have a link or button that does not imply the user has control; 4.4% hide their dialogue from EU IP addresses. TrustArc dialogues tend to define essential cookies for which there is no opt-out option. This, in combination with hiding dialogues from EU users, results from the product being tailored to the CCPA.

Finally, we estimate that about 8% of websites use CMPs for their APIs only and design custom consent dialogues themselves. This form of publisher customization presents a very practical problem: while these websites collect a standardized form of consent, each website does so in their own unique way, which may or may not comply with local legislation. As CMPs share consent across websites [60], this unreliable consent signal will then be re-used by other websites and third parties.

5.4.2 *Measuring Third Party Vendors*

The next two items of interest concern the purposes and lawful basis claimed by vendors for processing personal data. Using conventional methods, estimating how third-parties use personal data would require accessing and processing the privacy policy of each, which could be costly if repeated for longitudinal insights. In contrast, the IAB’s standard allows us to measure this longitudinally for vendors on the Global Vendor List (GVL). In fact, the organization managing the GVL switched to weekly updates so we can detect all changes.

Figure 5.7 speaks to **I4**. It shows that both the size of the number of vendors and the reported purposes in the IAB’s Global Vendor List have grown

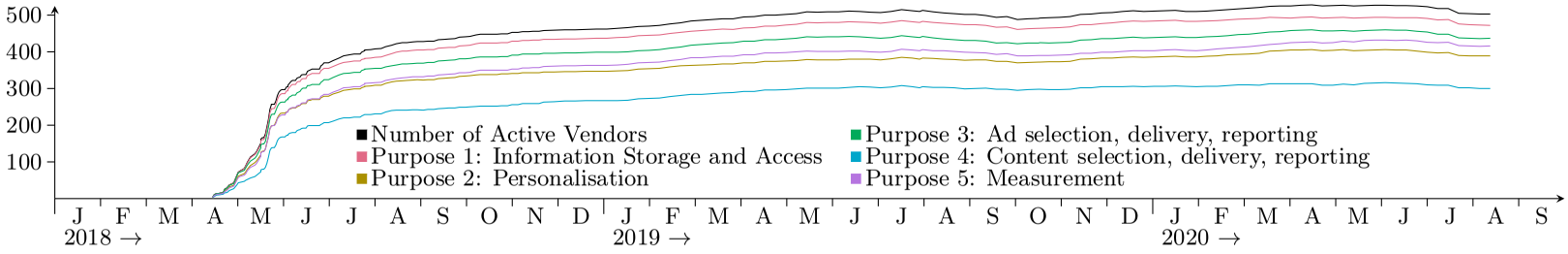


Figure 5.7: Reported purposes in the IAB's Global Vendor List

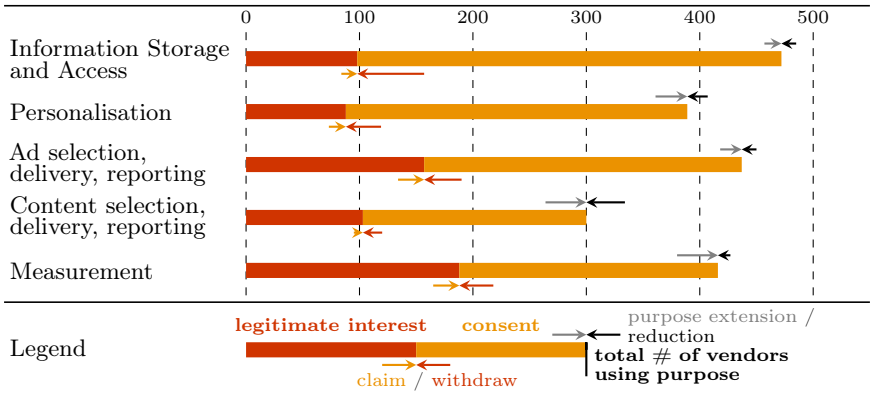


Figure 5.8: Purposes recorded in the IAB Global Vendor List

over time, with a sharp spike as GDPR came into effect. The first purpose, which allows vendors to collect and access personal data, is always the most popular. In Figure 5.7, it is difficult to track which movements are due to firms joining and which are due to an existing coalition member changing.

The changes made by existing members are summarized in Figure 5.8. This shows the surprising result that on net more vendors are now obtaining consent for purposes they used to claim as a legitimate interest than the other way round, which speaks to **I5**. This suggests that as time has passed, vendors on the GVL are obtaining more consent. The most activity regarding these changes took place around GDPR coming into effect, followed by another bout of activity in March and April 2020, possibly as vendors saw how GDPR was being enforced.

5.4.3 *Measuring the User-Interface*

Our results conclude with some findings regarding time costs related to consent dialogues. Our first item of interest here is the time it takes to send consent signals to multiple vendors (**I6**). We repeatedly measured the user's waiting time when they opt-out on a consent dialog provided by TrustArc and report the median numbers here. Figure 5.9 shows the opt-out process, which takes at least 7 clicks and 34s to complete (not including user interaction). This delay results from sending opt-out requests to multiple third parties and additional JavaScript timeouts. Compared to accepting cookies, opting out causes an additional 279 HTTP(S) requests to 25 domains, which amounts to an additional 1.2 MB / 5.8 MB of data transfer (compressed / uncompressed). Thus in 12% of the websites embedding TrustArc (see Section 5.4.1), opting out is associated with a significant time and network cost for the user.

Second, we measured how the dialog interaction time varies depending on which privacy preferences are expressed (**I7**). Instead of using an artificial dialog design, we conducted a randomized experiment using Quantcast's real consent dialog in two different configurations further described in Section 5.3.2. In short, the first configuration included a direct reject button which was

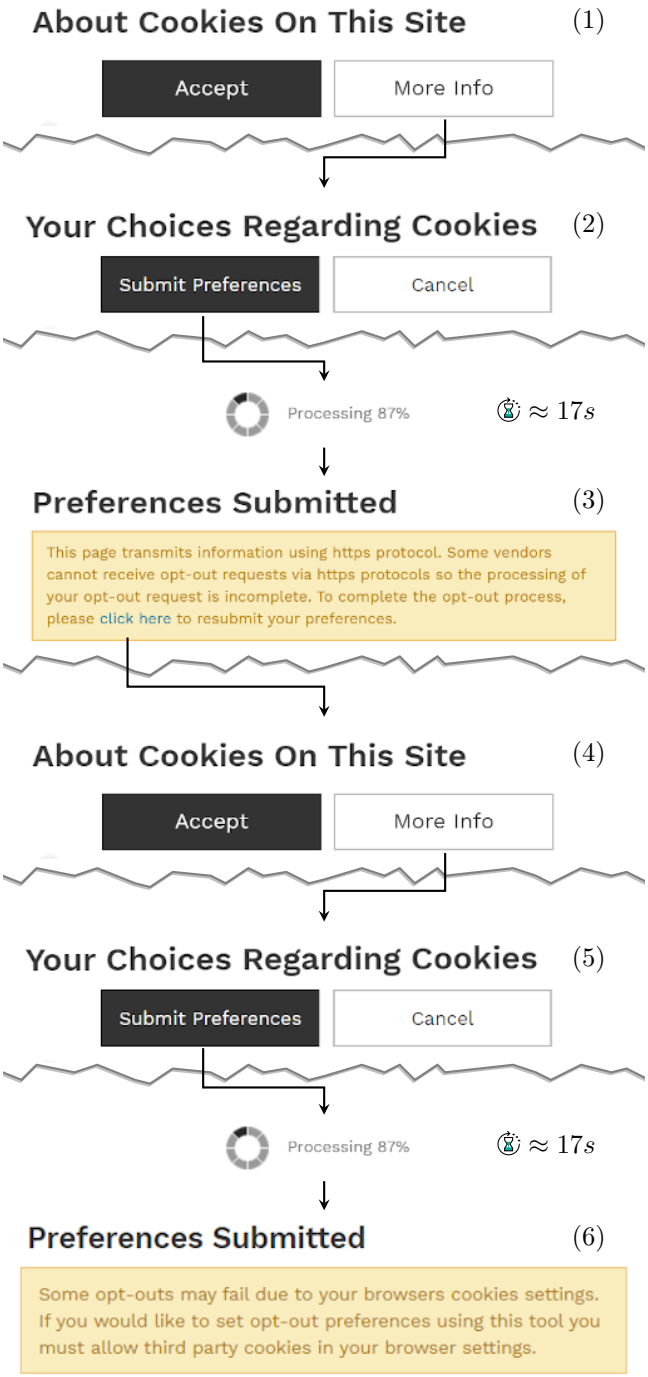


Figure 5.9: Training users to accept: Opting out on `forbes.com` takes at least 34 seconds (and seven clicks). Accepting cookies closes the dialog immediately.

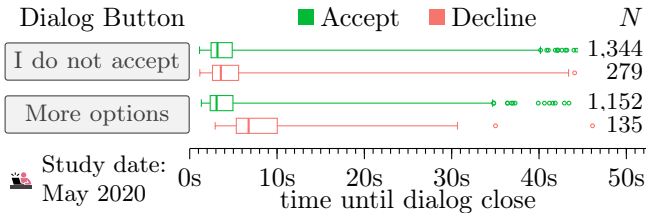


Figure 5.10: Randomized experiment with real CMP dialogs: depending on the dialog design, denying consent may take significantly longer than giving.

replaced with a "More Options" button in the second one (see Figures 5.11–5.13). Section 5.4.1 showed that the first and second option were respectively used by 55% and 45% of websites embedding Quantcast dialogues. We exclude users who made no decision within the first three minutes after page load. In total, consent dialogs were shown to 2910 visitors from the EU (as per Quantcast's default configuration).

Our results are summarized in Figure 5.10: If Quantcast's dialog with a direct reject button is shown, it took the median user 3.2s to accept and 3.6s to deny consent. This difference is small but already statistically significant using a nonparametric test that is robust to skewed distibutions (Mann–Whitney $U(N_{\text{accept}} = 1344, N_{\text{reject}} = 279) = 166582$, $z = -2.93$, $p < 0.01$). If no direct reject button is shown, the median time it takes users to deny consent doubles to 6.7 seconds, which is highly significant ($U(N_{\text{accept}} = 1152, N_{\text{reject}} = 135) = 30494$, $z = -11.57$, $p < 0.001$). Additionally, the consent rate increases from 83% to 90%. In summary, we find that depending on the dialog design, the interaction time increases greatly for users who intend to opt out.

5.5 DISCUSSION

Section 5.5.1 discusses measurement issues like sampling and generalizing. Section 5.5.2 discusses the prevalence, significance, and future of consent management provision.

5.5.1 *Methodological Implications*

SOCIAL MEDIA SAMPLING Sampling URLs from social media posts is a novel approach through which we captured 161 million web pages from 4.2 million unique domains over a period of 2.5 years. This significantly exceeds the sample size and windows used in related work (see Figure 5.1). Building on recent approaches [55], subsite sampling is more tolerant to the many idiosyncrasies regarding how CMPs are embedded in the wild. At the same time, this sample is influenced by the social media websites' content filtering policies and–more importantly–heavily skewed towards the 'attention economy'. Such websites tend to be funded by collecting personal data, for which consent

needs to be obtained. This bias is useful as we are more likely to sample websites that include CMPs.

We complement our social media crawling with a more traditional approach using the Tranco toplist. This means the proportions we estimate in Figures 5.5 and 5.6 are not affected by the social media sampling bias. However, top-lists are not representative of a meaningful population either, such as total web-page views or distinct sites visited by users. Given that both bottom-up sampling from social media posts and top-down sampling from toplists over-samples a certain population [49] with no ground-truth to adjust for it, using both approaches seems a defensible way forward.

WEB PRIVACY MEASUREMENTS The notion that a web-page has a single set of observer-independent privacy features is dead [58]. We demonstrated that CMP adoption is influenced by local legislation and measurement results depend on vantage point (see Figure 5.1). Future studies should consider this and explain the implications for generalizing findings if only one vantage point is used.

Similarly, the occurrence of CMPs varies greatly depending on the toplist size (see Figure 5.5). From 4% in the Top 100, it reaches 13% in the Top 1k, and then falls in the long-tail down to 1.51% for the Top 1M. These stark differences emphasize the importance of both sample size and choice of toplist from which it is drawn.

Web scraping can exploit common code structure across websites embedding CMPs. Such research designs can be scaled across the long tail of website popularity, which complements the qualitative analysis of tech giants [18]. However, it is not clear how such results generalize beyond websites employing CMPs. Similarly, we do not know how our results, based on six of the most popular CMPs, apply to niche CMPs[6] or websites self-implementing the TCF framework.

MEASURING AD-TECH BEHAVIOR Given frameworks such as the TCF, the legal basis for third-party vendors can now be publicly queried and measured over time (see Figures 5.7 and 5.8) whereas previously this information was stored on corporate networks. However, these frameworks only provide self-reported privacy policy. It remains a challenge to audit compliance.

5.5.2 *Privacy Implications*

PREVALENCE We observed that CMPs are embedded in ever more websites over time and that privacy laws coming into effect caused spikes in adoption. The few times the GDPR was enforced had little observable effect (see Figure 5.6), although this could change if sanctions increase in frequency or significance. There is further churn between CMPs with Cookiebot functioning

[6] Examples include Kochava, Adzerk CMP, and PreferenceManager.

as a 'gateway CMP' that many websites adopt before migrating onto other CMPs (see Figure 5.4).

SIGNIFICANCE CMPs are standardizing privacy communications. The resulting legal terms, dialogue interface, and protocol for communicating with vendors should be seen as a de-facto standard, at least among that CMP's customers. Such standards were developed by self-interested private companies and not in the open bodies like the IETF or W3C, which raises questions about the politics of standards [26]. More positively, the consistent web interfaces provided by CMPs help researchers discover possible privacy violations at scale [39, 32], which mirrors researchers auditing compliance to credit card security standards [46, 31].

Beyond technical standards, CMPs can also influence social norms around privacy by herding websites. This can be seen in the linguistic shift from cookies to *scripts* that was only observed in 5.5% of the websites embedding OneTrust. This is likely a strategic move to escape the negative associations of cookies [54]. Herding may also strengthen the widely documented habituation effect in both privacy [5, 59, 24] and security notices [12].

Compliance with privacy laws drives CMP adoption, as evidenced by the spikes after the laws come into effect, and yet liability for violations is an open question. Quantcast maintain that "with great customizability comes great responsibility", which suggests they believe websites are liable for using terms like "whatever" as an affirmative signal of consent. Yet Quantcast offer dialogue functionality in which accepting takes 1-click while rejecting takes multiple, which is adopted by 45% of their customers, despite the French regulator's guidance against this practice [10].

Buttons allowing 1-click rejection are even rarer among websites embedding TrustArc (7%) and OneTrust (2.4%). The CMPs may know something its clients do not given `trustarc.com` implements an instant, 1-click reject all button. Disentangling whether these differences are driven by CMP business practices or pre-existing customer characteristics (e.g jurisdiction) can help prioritize regulatory interventions. The role of intermediaries in (not) preventing abuse is an endearing lesson from information security economics [35, 52, 8], why would privacy economics be any different?

The specter of liability looms over vendors claiming a legitimate interest rather than obtaining consent [33]. For every purpose in the TCF, at least a fifth of the vendors claim they do not need to collect consent to process personal data (see Figure 5.8). More generally, one might ask why websites agree to collect consent for all of the Global Vendor List given there is no observed benefit to doing so [60].

THE FUTURE OF CONSENT MANAGEMENT If trends during the formation of the ecosystem continue, Figure 5.4 suggests that certain CMPs (Quantcast, OneTrust) will win market share from the others. A theoretical model predicts that sharing consent between the CMP's customers will create

winner takes all dynamics leading to one global coalition [60]. In reality, jurisdictional boundaries will likely lead to multiple distinct coalitions given Quantcast and OneTrust appear to be establishing dominance in the EU+UK and the US respectively. However, users do not respect such jurisdictions. This will likely exacerbate the extent to which the web differs based on where the user appears to be located, which we observed at multiple points in this study.

The rise of CMPs should be seen as part of a wider process by which legal compliance shapes the internet. Liability for content shared on technology platforms provides another example [15] in light of a May 2020 executive order in the US. This represents a departure from utopian views of the Internet as a libertarian paradise [3]. One might begin to consider a *compliance layer* of the internet driven by the content and privacy policies of private firms as influenced by national laws. Before regulators demand measurements as evidence, the community should reflect on how to support auditing at scale, evidential standards, and surrounding ethical issues.

5.6 RELATED WORK

Returning to the piping metaphor of Figure 5.2, consent flows from a user's privacy preferences through a consent dialogue to the recipient of the consent signal and then on to third-parties. This section identifies related work at each interface, though none of the studies make measurements at as many interfaces as we do.

Qualitative research exploring privacy preferences of users informs internet design by, for example, identifying disparities between what users want and what happens online [4, 23, 40] or by highlighting the business value of obtaining explicit consent [61].

At the user-interface, lab experiments have consistently shown users can be shifted towards providing consent by changing framing [5, 2] and design choices, such as default settings [28, 30] and positioning [56]. Nouwens et al. [39] scrape post-GDPR UK websites to identify popular design choices and show that common practices like not having fine-grained controls on the first page increases propensity to consent. Our controlled experiment with real CMP dialogs on a public website complemented this body of work by showing users incur differing time costs based on the privacy preferences they express, highlighting how this punishes privacy aware users.

The next point of the consent flow concerns how consent dialogues interact with websites. Around 50% of the websites in [39] do not offer a 1-click opt out, which is confirmed by our samples of Quantcast websites. A dialogue or cookie banner may not even be shown. Degeling et al. [11] showed that 62 % of sampled European websites displayed cookie prompts right after GDPR came into effect in May 2018, up from 46 % in January 2018. However, these effects are not limited to Europe as websites in the US "approach cookie regulations similarly to the EU" [48], though this is not true of Chinese websites.

Turning to third-parties, research has predominantly focused on the extent of third-party tracking rather than how third-parties obtain consent (the final part of the consent flow). Iordanou et al. [22] introduce a methodology for measuring tracking at scale and show that the majority of tracking flows across European borders but, surprisingly, remains within the EU. Sørensen and Kosta [50] do not establish any change in the number of third-party trackers before and after GDPR, although they show that third-party tracking is more prevalent in private websites than public. Even after GDPR, Sanchez-Rola et al. [48] show that 90 % of sampled websites use cookies that could be used to identify users. Such results are hard to evaluate without more context. For example, a website needs to identify users who have not consented in order to not repeatedly present consent dialogues, which would violate the California Consumer Privacy Act.

Basing measurement on the TCF standard provides a way forward, Matte et al. [32] analyze sites using the TCF and find disparities between which preferences were communicated and which were stored as global cookies, which is more reliable evidence of a privacy violation. For example, 12 % of websites send the consent signal before the user even makes a choice and some even record the user's consent after an explicit opt-out. In a different study, the same authors argue that the purposes in the TCF are not specific or explicit enough "to be used as legally-compliant ones" [33] and measure which vendors claim these as a legitimate interest.

Finally, a theoretical work [60] considers the economic implications of CMPs forming 'consent coalitions' in which consent is shared across websites and vendors. Our measurements contradict their theoretical prediction about a 'global coalition', which does not exist at present. The market will, however, further mature and our longitudinal results suggest a trend towards dominant CMPs in particular jurisdictions.

Considering our contribution to each aspect of online privacy in isolation obscures how our measurement approach allowed us to make longitudinal measurements across the entire consent ecosystem. Similar ecosystem wide measurements include those of: the advertising industry [14, 45, 43, 42]; online gaming [41]; VPN services [25, 21]; web communities [63, 62]; and web porn [57]. All of these studies, including ours, blend technical measurements with considerations around the economic and social factors influencing the agents in the ecosystem. Such studies provide a rigorous, empirical basis for how social scientists theorize about the impact of the Internet.

5.7 CONCLUSION

Recent years have seen the formation of a consent ecosystem through which websites and third-party vendors establish a legal basis for business models based on personal data. Our longitudinal approach tracks the rise of CMPs from less than 1% of the Tranco 10k toplist in February 2018 to almost 10% in September 2020 and we show that privacy laws (GDPR and CCPA) coming

into effect caused spikes in adoption. We document inter-firm competition by which certain CMPs (e.g Cookiebot) bleed customers while others slowly establish dominance in a specific jurisdiction, such as Quantcast in the EU+UK or OneTrust in the US. This increasing market dominance allows private actors (often tied to the Ad-tech industry) to standardize the terms user consent to, the user-interface through which they do it, and also how it is shared with third-parties.

Although increasing market power is worrying, the same standardization opens up novel measurements opportunities. We tracked how third-party vendors justified their data processing activities, capturing changes over time like the shift towards obtaining consent. Similarly, we showed how the consent dialogues offered by CMPs impose a time cost on privacy aware users. These exact dialogues are used by the CMP's customers, which improved the ecological validity of our real-user study. More generally, regulators could exploit the structure provided by CMPs to audit privacy practices at scale.

ACKNOWLEDGEMENTS

We would like to thank Aldo Cortesi for his continuous support and the generous access to the Netograph API and capturing technology. We thank Tobias Kupek for his help with preparing figures. This work was co-funded by Archimedes Privatstiftung, Innsbruck. The second author is funded by the European Commission's call H2020-MSCA-IF-2019 under grant number 894700.

REFERENCES

[1] Gunes Acar, Christian Eubank, Steven Englehardt, Marc Juarez, Arvind Narayanan, and Claudia Diaz. 2014. The Web Never Forgets: Persistent Tracking Mechanisms in the Wild. In *Proceedings of the 2014 ACM SIGSAC Conference on Computer and Communications Security (CCS '14)*. ACM, 674–689. https://doi.org/10.1145/2660267.2660347

[2] Idris Adjerid, Alessandro Acquisti, Laura Brandimarte, and George Loewenstein. 2013. Sleights of Privacy: Framing, Disclosures, and the Limits of Transparency. In *Proceedings of the Ninth Symposium on Usable Privacy and Security (SOUPS '13)*. ACM, Article 9, 11 pages. https://doi.org/10.1145/2501604.2501613

[3] John Perry Barlow. 1996. A Declaration of the Independence of Cyberspace.

[4] Bettina Berendt, Oliver Günther, and Sarah Spiekermann. 2005. Privacy in E-Commerce: Stated Preferences vs. Actual Behavior. *Commun. ACM* 48, 4 (April 2005), 101–106. https://doi.org/10.1145/1053291.1053295

[5] Rainer Böhme and Stefan Köpsell. 2010. Trained to Accept? A Field Experiment on Consent Dialogs. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '10)*. ACM, 2403–2406. https://doi.org/10.1145/1753326.1753689

[6] Joseph Bonneau and Sören Preibusch. 2009. The Privacy Jungle: On the Market for Data Protection in Social Networks. In *8th Annual Workshop on the Economics of Information Security, WEIS*. https://doi.org/10.1007/978-1-4419-6967-5_8

[7] Aaron Ceross and Andrew Simpson. 2018. Rethinking the Proposition of Privacy Engineering. In *Proceedings of the New Security Paradigms Workshop (NSPW '18)*. ACM, 89–102. https://doi.org/10.1145/3285002.3285006

[8] Richard Clayton, Tyler Moore, and Nicolas Christin. 2015. Concentrating Correctly on Cybercrime Concentration. In *14th Annual Workshop on the Economics of Information Security, WEIS*.

[9] Lorrie Faith Cranor. 2003. P3P: Making Privacy Policies More Useful. *IEEE Security and Privacy* 1, 6 (2003), 50–55. https://doi.org/10.1109/MSECP.2003.1253568

[10] Commission Nationale de l'Informatique et des Libertés (CNIL). 2019. *Guidelines on cookies and tracking devices*. https://www.cnil.fr/en/cookies-and-other-tracking-devices-cnil-publishes-new-guidelines

[11] Martin Degeling, Christine Utz, Christopher Lentzsch, Henry Hosseini, Florian Schaub, and Thorsten Holz. 2019. We Value Your Privacy ... Now Take Some Cookies: Measuring the GDPR's Impact on Web Privacy. In *26th Annual Network and Distributed System Security Symposium (NDSS '19)*. The Internet Society. https://doi.org/10.14722/ndss.2019.23378

[12] Serge Egelman, Lorrie Faith Cranor, and Jason Hong. 2008. You've Been Warned: An Empirical Study of the Effectiveness of Web Browser Phishing Warnings. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '08)*. ACM, 1065–1074. https://doi.org/10.1145/1357054.1357219

[13] Mozilla Foundation. 2007–2020. *Public Suffix List*. https://publicsuffix.org/

[14] Phillipa Gill, Vijay Erramilli, Augustin Chaintreau, Balachander Krishnamurthy, Konstantina Papagiannaki, and Pablo Rodriguez. 2013. Follow the Money: Understanding Economics of Online Aggregation and Advertising. In *Proceedings of the 2013 Conference on Internet Measurement Conference (IMC '13)*. ACM, 141–148. https://doi.org/10.1145/2504730.2504768

[15] Tarleton Gillespie. 2010. The politics of 'platforms'. *New Media & Society* 12, 3 (2010), 347–364. https://doi.org/10.1177/1461444809342738

[16] Hana Habib, Yixin Zou, Aditi Jannu, Neha Sridhar, Chelse Swoopes, Alessandro Acquisti, Lorrie Faith Cranor, Norman M. Sadeh, and Florian Schaub. 2019. An Empirical Analysis of Data Deletion and Opt-Out Choices on 150 Websites. In *Fifteenth Symposium on Usable Privacy and Security (SOUPS '19)*. USENIX. https://www.usenix.org/conference/soups2019/presentation/habib

[17] Elizabeth Liz Harding, Jarno J Vanto, Reece Clark, L Hannah Ji, and Sara C Ainsworth. 2019. Understanding the scope and impact of the California Consumer Privacy Act of 2018. *Journal of Data Protection & Privacy* 2, 3 (2019), 234–253.

[18] Soheil Human and Florian Cech. 2021. A Human-centric Perspective on Digital Consenting: The Case of GAFAM. In *Human Centred Intelligent Systems*. Springer, 139–159. https://doi.org/10.1007/978-981-15-5784-2_12

[19] IAB Europe. 2020. *CMP List*. https://iabeurope.eu/cmp-list/

[20] IAB Europe. 2020. *What is the Transparency and Consent Framework (TCF)?* https://iabeurope.eu/transparency-consent-framework/

[21] Muhammad Ikram, Narseo Vallina-Rodriguez, Suranga Seneviratne, Mohamed Ali Kaafar, and Vern Paxson. 2016. An Analysis of the Privacy and

Security Risks of Android VPN Permission-Enabled Apps. In *Proceedings of the 2016 Internet Measurement Conference (IMC '16)*. ACM, 349–364. https://doi.org/10.1145/2987443.2987471

[22] Costas Iordanou, Georgios Smaragdakis, Ingmar Poese, and Nikolaos Laoutaris. 2018. Tracing Cross Border Web Tracking. In *Proceedings of the Internet Measurement Conference 2018 (IMC '18)*. ACM, 329–342. https://doi.org/10.1145/3278532.3278561

[23] Ruogu Kang, Laura Dabbish, Nathaniel Fruchter, and Sara B. Kiesler. 2015. "My Data Just Goes Everywhere:" User Mental Models of the Internet and Implications for Privacy and Security. In *Eleventh Symposium On Usable Privacy and Security (SOUPS '15)*. USENIX, 39–52. https://www.usenix.org/conference/soups2015/proceedings/presentation/kang

[24] Farzaneh Karegar, John Sören Pettersson, and Simone Fischer-Hübner. 2020. The Dilemma of User Engagement in Privacy Notices: Effects of Interaction Modes and Habituation on User Attention. *ACM Transactions on Privacy and Security* 23, 1 (2020), 5:1–5:38. https://doi.org/10.1145/3372296

[25] Mohammad Taha Khan, Joe DeBlasio, Geoffrey M. Voelker, Alex C. Snoeren, Chris Kanich, and Narseo Vallina-Rodriguez. 2018. An Empirical Analysis of the Commercial VPN Ecosystem. In *Proceedings of the Internet Measurement Conference 2018 (IMC '18)*. ACM, 443–456. https://doi.org/10.1145/3278532.3278570

[26] David M. Kristol. 2001. HTTP Cookies: Standards, privacy, and politics. *ACM Transactions on Internet Technology* 1, 2 (2001), 151–198. https://doi.org/10.1145/502152.502153

[27] Ponnurangam Kumaraguru, Lorrie Cranor, Jorge Lobo, and Seraphin Calo. 2007. A Survey of Privacy Policy Languages. In *Workshop on Usable IT Security Management: 3rd Symposium on Usable Privacy and Security, ACM (USM '07)*.

[28] Yee-Lin Lai and Kai-Lung Hui. 2006. Internet Opt-in and Opt-out: Investigating the Roles of Frames, Defaults and Privacy Concerns. In *Proceedings of the 2006 ACM SIGMIS CPR Conference on Computer Personnel Research (SIGMIS CPR '06)*. ACM, 253–263. https://doi.org/10.1145/1125170.1125230

[29] Pierre Laperdrix, Nataliia Bielova, Benoit Baudry, and Gildas Avoine. 2020. Browser Fingerprinting: A Survey. *ACM Transactions on the Web* 14, 2, Article 8 (April 2020), 33 pages. https://doi.org/10.1145/3386040

[30] Dominique Machuletz and Rainer Böhme. 2020. Multiple Purposes, Multiple Problems: A User Study of Consent Dialogs after GDPR.

Proceedings on Privacy Enhancing Technologies 2, 481–498. https://doi.org/10.2478/popets-2020-0037

[31] Samin Yaseer Mahmud, Akhil Acharya, Benjamin Andow, William Enck, and Bradley Reaves. 2020. Cardpliance: PCI DSS Compliance of Android Applications. In *29th USENIX Security Symposium (USENIX '20)*. 1517–1533. https://www.usenix.org/conference/usenixsecurity20/presentation/mahmud

[32] Célestin Matte, Nataliia Bielova, and Cristiana Santos. 2020. Do Cookie Banners Respect my Choice? Measuring Legal Compliance of Banners from IAB Europe's Transparency and Consent Framework. In *2020 IEEE Symposium on Security and Privacy*. IEEE, 791–809. https://doi.org/10.1109/SP40000.2020.00076

[33] Célestin Matte, Cristiana Santos, and Nataliia Bielova. 2020. Purposes in IAB Europe's TCF: which legal basis and how are they used by advertisers?. In *Annual Privacy Forum (APF 2020)*.

[34] Jonathan R. Mayer and John C. Mitchell. 2012. Third-Party Web Tracking: Policy and Technology. In *2012 IEEE Symposium on Security and Privacy*. IEEE, 413–427. https://doi.org/10.1109/SP.2012.47

[35] Damon McCoy, Hitesh Dharmdasani, Christian Kreibich, Geoffrey M. Voelker, and Stefan Savage. 2012. Priceless: The Role of Payments in Abuse-Advertised Goods. In *Proceedings of the 2012 ACM Conference on Computer and Communications Security (CCS '12)*. ACM, 845–856. https://doi.org/10.1145/2382196.2382285

[36] Aleecia M McDonald and Lorrie Faith Cranor. 2008. The cost of reading privacy policies. *Journal of Law and Policy for the Information Society* 4 (2008), 543.

[37] Lynette I. Millett, Batya Friedman, and Edward Felten. 2001. Cookies and Web Browser Design: Toward Realizing Informed Consent Online. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '01)*. ACM, 46–52. https://doi.org/10.1145/365024.365034

[38] Nick Nikiforakis, Alexandros Kapravelos, Wouter Joosen, Christopher Kruegel, Frank Piessens, and Giovanni Vigna. 2013. Cookieless Monster: Exploring the Ecosystem of Web-Based Device Fingerprinting. In *2013 IEEE Symposium on Security and Privacy*. IEEE, 541–555. https://doi.org/10.1109/SP.2013.43

[39] Midas Nouwens, Ilaria Liccardi, Michael Veale, David Karger, and Lalana Kagal. 2020. Dark Patterns after the GDPR: Scraping Consent Pop-Ups and Demonstrating Their Influence. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems (CHI '20)*. ACM, 1–13. https://doi.org/10.1145/3313831.3376321

[40] Judith S. Olson, Jonathan Grudin, and Eric Horvitz. 2005. A Study of Preferences for Sharing and Privacy. In *CHI '05 Extended Abstracts on Human Factors in Computing Systems*. ACM, 1985–1988. https://doi.org/10.1145/1056808.1057073

[41] Mark O'Neill, Elham Vaziripour, Justin Wu, and Daniel Zappala. 2016. Condensing Steam: Distilling the Diversity of Gamer Behavior. In *Proceedings of the 2016 Internet Measurement Conference (IMC '16)*. ACM, 81–95. https://doi.org/10.1145/2987443.2987489

[42] Michalis Pachilakis, Panagiotis Papadopoulos, Evangelos P. Markatos, and Nicolas Kourtellis. 2019. No More Chasing Waterfalls: A Measurement Study of the Header Bidding Ad-Ecosystem. In *Proceedings of the Internet Measurement Conference (IMC '19)*. ACM, 280–293. https://doi.org/10.1145/3355369.3355582

[43] Panagiotis Papadopoulos, Nicolas Kourtellis, Pablo Rodriguez Rodriguez, and Nikolaos Laoutaris. 2017. If You Are Not Paying for It, You Are the Product: How Much Do Advertisers Pay to Reach You?. In *Proceedings of the 2017 Internet Measurement Conference (IMC '17)*. ACM, 142–156. https://doi.org/10.1145/3131365.3131397

[44] Victor Le Pochat, Tom van Goethem, Samaneh Tajalizadehkhoob, Maciej Korczynski, and Wouter Joosen. 2019. Tranco: A Research-Oriented Top Sites Ranking Hardened Against Manipulation. In *26th Annual Network and Distributed System Security Symposium (NDSS '19)*. The Internet Society. https://doi.org/10.14722/ndss.2019.23386

[45] Enric Pujol, Oliver Hohlfeld, and Anja Feldmann. 2015. Annoyed Users: Ads and Ad-Block Usage in the Wild. In *Proceedings of the 2015 Internet Measurement Conference (IMC '15)*. ACM, 93–106. https://doi.org/10.1145/2815675.2815705

[46] Sazzadur Rahaman, Gang Wang, and Danfeng (Daphne) Yao. 2019. Security Certification in Payment Card Industry: Testbeds, Measurements, and Recommendations. In *Proceedings of the 2019 ACM SIGSAC Conference on Computer and Communications Security (CCS '19)*. ACM, 481–498. https://doi.org/10.1145/3319535.3363195

[47] Hana Ross. 2017. Data subject consent: How will the General Data Protection Regulation affect this? *Journal of Data Protection & Privacy* 1, 2 (2017), 146–155.

[48] Iskander Sanchez-Rola, Matteo Dell'Amico, Platon Kotzias, Davide Balzarotti, Leyla Bilge, Pierre-Antoine Vervier, and Igor Santos. 2019. Can I Opt Out Yet? GDPR and the Global Illusion of Cookie Control. In *Proceedings of the 2019 ACM Asia Conference on Computer and Communications Security (Asia CCS '19)*. ACM, 340–351. https://doi.org/10.1145/3321705.3329806

[49] Quirin Scheitle, Oliver Hohlfeld, Julien Gamba, Jonas Jelten, Torsten Zimmermann, Stephen D. Strowes, and Narseo Vallina-Rodriguez. 2018. A Long Way to the Top: Significance, Structure, and Stability of Internet Top Lists. In *Proceedings of the Internet Measurement Conference 2018 (IMC '18)*. ACM, 478–493. https://doi.org/10.1145/3278532.3278574

[50] Jannick Sørensen and Sokol Kosta. 2019. Before and After GDPR: The Changes in Third Party Presence at Public and Private European Websites. In *The World Wide Web Conference (WWW '19)*. ACM, 1590–1600. https://doi.org/10.1145/3308558.3313524

[51] Sarah Spiekermann, Alessandro Acquisti, Rainer Böhme, and Kai Lung Hui. 2015. The challenges of personal data markets and privacy. *Electronic Markets* 25, 2 (2015), 161–167. https://doi.org/10.1007/s12525-015-0191-0

[52] Samaneh Tajalizadehkhoob, Tom Van Goethem, Maciej Korczyński, Arman Noroozian, Rainer Böhme, Tyler Moore, Wouter Joosen, and Michel van Eeten. 2017. Herding Vulnerable Cats: A Statistical Approach to Disentangle Joint Responsibility for Web Security in Shared Hosting. In *Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security (CCS '17)*. ACM, 553–567. https://doi.org/10.1145/3133956.3133971

[53] Martino Trevisan, Stefano Traverso, Eleonora Bassi, and Marco Mellia. 2019. 4 Years of EU Cookie Law: Results and Lessons Learned. *Proceedings on Privacy Enhancing Technologies* 2019, 2 (2019), 126–145. https://doi.org/10.2478/popets-2019-0023

[54] Blase Ur, Pedro Giovanni Leon, Lorrie Faith Cranor, Richard Shay, and Yang Wang. 2012. Smart, Useful, Scary, Creepy: Perceptions of Online Behavioral Advertising. In *Proceedings of the Eighth Symposium on Usable Privacy and Security (SOUPS '12)*. ACM, Article 4, 15 pages. https://doi.org/10.1145/2335356.2335362

[55] Tobias Urban, Martin Degeling, Thorsten Holz, and Norbert Pohlmann. 2020. Beyond the Front Page: Measuring Third Party Dynamics in the Field. In *Proceedings of The Web Conference 2020 (WWW '20)*. ACM, 1275–1286. https://doi.org/10.1145/3366423.3380203

[56] Christine Utz, Martin Degeling, Sascha Fahl, Florian Schaub, and Thorsten Holz. 2019. (Un)informed Consent: Studying GDPR Consent Notices in the Field. In *Proceedings of the 2019 ACM SIGSAC Conference on Computer and Communications Security (CCS '19)*. ACM, 973–990. https://doi.org/10.1145/3319535.3354212

[57] Pelayo Vallina, Álvaro Feal, Julien Gamba, Narseo Vallina-Rodriguez, and Antonio Fernández Anta. 2019. Tales from the Porn: A Comprehensive

Privacy Analysis of the Web Porn Ecosystem. In *Proceedings of the Internet Measurement Conference (IMC '19)*. ACM, 245–258. https://doi.org/10.1145/3355369.3355583

[58] Rob Van Eijk, Hadi Asghari, Philipp Winter, and Arvind Narayanan. 2019. The Impact of User Location on Cookie Notices (Inside and Outside of the European Union). In *Workshop on Technology and Consumer Protection (ConPro'19)*.

[59] Anthony Vance, David Eargle, Jeffrey L. Jenkins, C. Brock Kirwan, and Bonnie Brinton Anderson. 2019. The Fog of Warnings: How Non-essential Notifications Blur with Security Warnings. In *Fifteenth Symposium on Usable Privacy and Security (SOUPS '19)*. USENIX. https://www.usenix.org/conference/soups2019/presentation/vance

[60] Daniel W Woods and Rainer Böhme. 2020. The Commodification of Consent. In *20th Annual Workshop on the Economics of Information Security, WEIS*.

[61] Scott A Wright and Guang-Xin Xie. 2019. Perceived Privacy Violation: Exploring the Malleability of Privacy Expectations. *Journal of Business Ethics* 156, 1 (2019), 123–140. https://doi.org/10.1007/s10551-017-3553-z

[62] Savvas Zannettou, Tristan Caulfield, Jeremy Blackburn, Emiliano De Cristofaro, Michael Sirivianos, Gianluca Stringhini, and Guillermo Suarez-Tangil. 2018. On the Origins of Memes by Means of Fringe Web Communities. In *Proceedings of the Internet Measurement Conference 2018 (IMC '18)*. ACM, 188–202. https://doi.org/10.1145/3278532.3278550

[63] Savvas Zannettou, Tristan Caulfield, Emiliano De Cristofaro, Nicolas Kourtelris, Ilias Leontiadis, Michael Sirivianos, Gianluca Stringhini, and Jeremy Blackburn. 2017. The Web Centipede: Understanding How Web Communities Influence Each Other through the Lens of Mainstream and Alternative News Sources. In *Proceedings of the 2017 Internet Measurement Conference (IMC '17)*. ACM, 405–417. https://doi.org/10.1145/3131365.3131390

APPENDIX

Purposes Definitions	
1	**Information storage and access:** The storage of information, or access to information that is already stored, on your device such as advertising identifiers, device identifiers, cookies, and similar technologies.
2	**Personalisation.** The collection and processing of information about your use of this service to subsequently personalise advertising and/or content for you in other contexts, such as on other websites or apps, over time.
3	**Ad selection, delivery, reporting.** The collection of information, and combination with previously collected information, to select and deliver advertisements for you, and to measure the delivery and effectiveness of such advertisements.
4	**Content selection, delivery, reporting.** The collection of information, and combination with previously collected information, to select and deliver content for you, and to measure the delivery and effectiveness of such content.
5	**Measurement.** The collection of information about your use of the content, and combination with previously collected information, used to measure, understand, and report on your usage of the service.
Careful readers may note that "information storage and access" is not a purpose for personal data processing in itself, but an artifact of the obligations imposed by Article 5(3) of the ePrivacy Directive.	

Feature Definitions	
1	**Offline data matching.** Combining data from offline sources that were initially collected in other contexts with data collected online in support of one or more purposes.
2	**Device linking.** Processing data to link multiple devices that belong to the same user in support of one or more purposes.
3	**Precise geographic location data.** Collecting and supporting precise geographic location data in support of one or more purposes.

Table 5.2: Purposes and features as defined in version 1 of the IAB's Trust and Consent Framework.

CMP	Unique Hostname
OneTrust	cdn.cookielaw.org
Quantcast	quantcast.mgr.consensu.org
TrustArc	consent.trustarc.com
Cookiebot	consent.cookiebot.com
LiveRamp	cmp.choice.faktor.io
Crownpeak	iabmap.evidon.com

Table 5.3: Hostnames used as an indicator for the presence of a CMP (see Section 5.3.2).

Location	US ☁	EU ☁	EU University		
User Agent	🇺🇸	🇺🇸	🇺🇸	🇬🇧	🇩🇪
Timing			⏱	⏱	⏱
OneTrust	263	306	344	339	342
Quantcast	151	192	222	220	221
TrustArc	102	110	170	168	168
Cookiebot	82	90	92	92	92
LiveRamp	6	6	10	10	10
Crownpeak	9	10	34	35	34
$\sum$	613	714	872	864	867
Coverage	70%	82%	100%	99%	99%

Table 5.4: Occurence of CMPs measured in January 2020. Comparing this to the May 2020 data in Table 5.1, we see that a growing share of websites adapt CMPs outside the EU, likely prompted by non-EU regulations such as CCPA.

Item of Interest	Vantage Point	Dataset
I1 CMP Adoption (by rank)	US/EU Cloud	Social media URLs from Tranco 1M
I2 CMP Adoption (over time)	US/EU Cloud	Social media URLs from Tranco 10k
I3 Publisher Customization	EU University	Tranco 10k front pages
I4 Collection Purposes	–	IAB Global Vendor List
I5 Legal Basis for Collection	–	IAB Global Vendor List
I6 Cost to Opt-Out	EU University	Measurements for `forbes.com`
I7 User Behavior	Visitors from EU countries	User study hosted on `mitmproxy.org`

Table 5.5: Overview of the vantage points and datasets used for each measurement (see Section 5.3.2). For the first two items of interest, each URL is randomly distributed to either a US or a EU cloud instance for crawling.

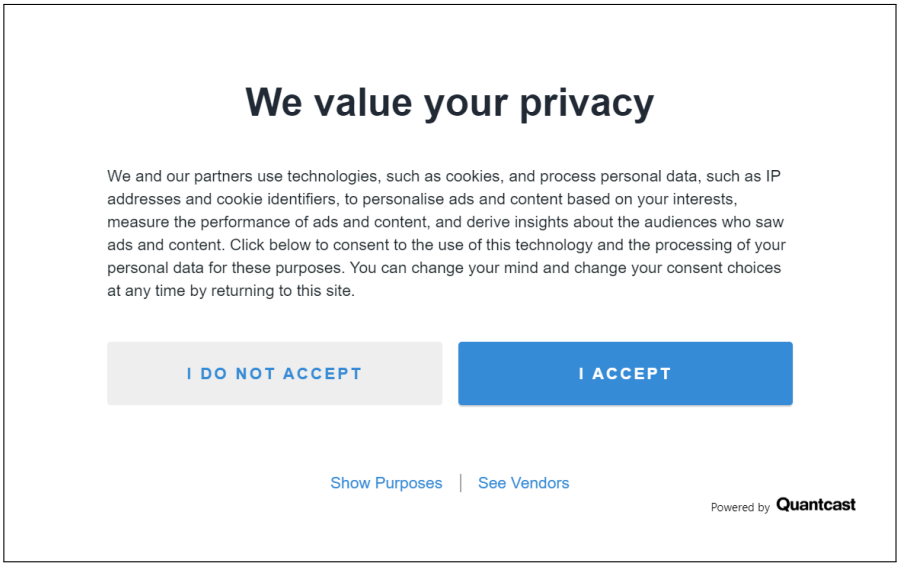


Figure 5.11: Default version of Quantcast's consent dialog. The dialog is shown as a modal popup with a dark-gray background covering the rest of the page.

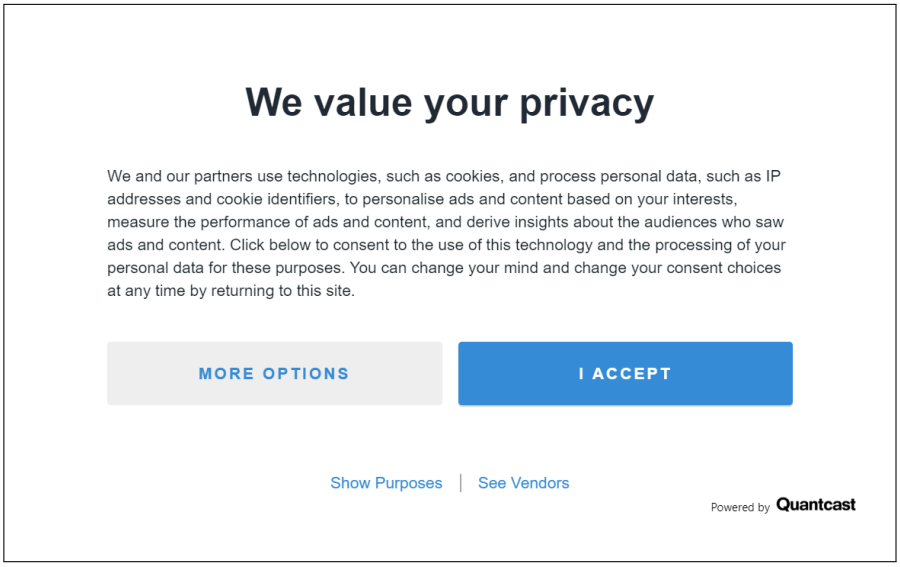


Figure 5.12: Quantcast's consent dialog without direct reject option.

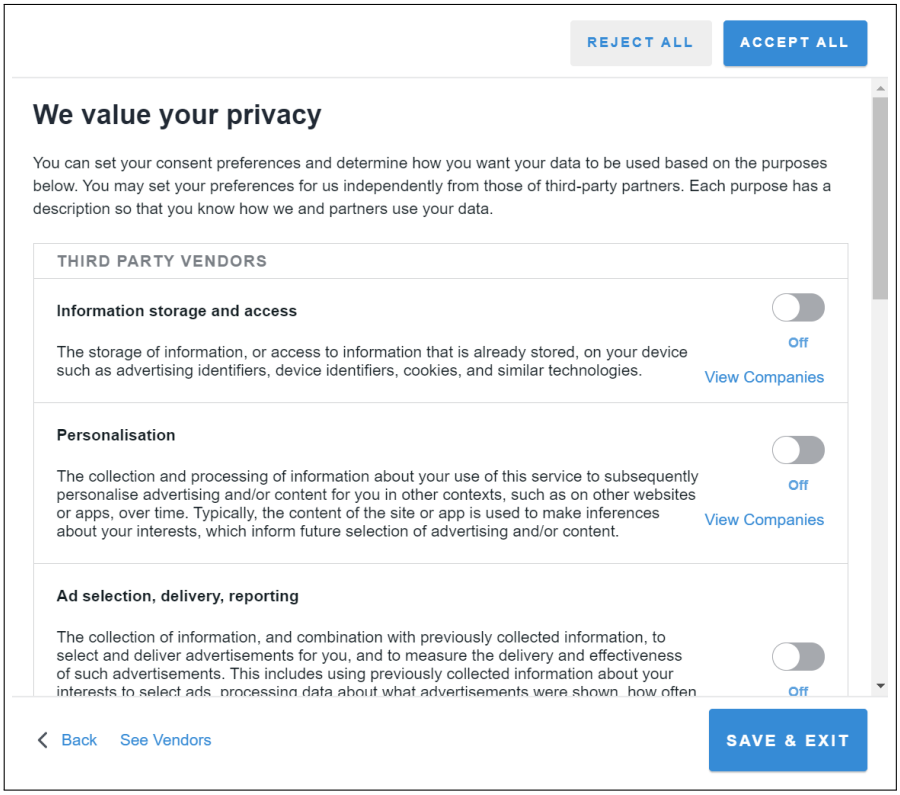


Figure 5.13: Dialog shown to users after they click “More Options”.

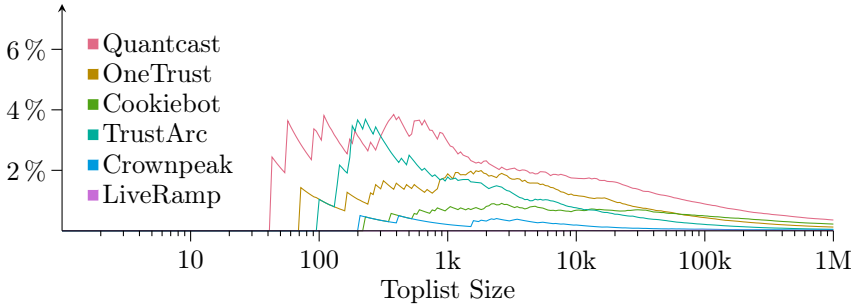


Figure 5.14: CMP Marketshare (January 2019).

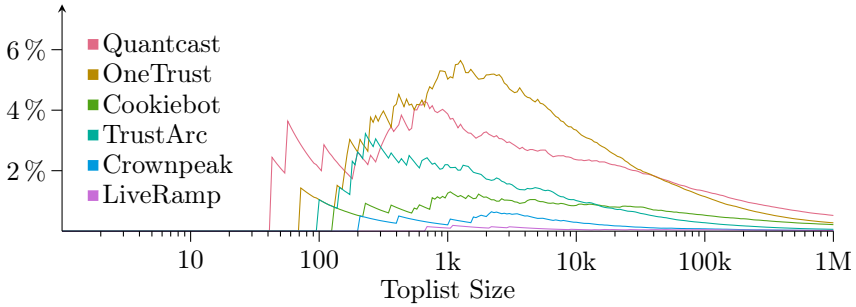


Figure 5.15: CMP Marketshare (January 2020).

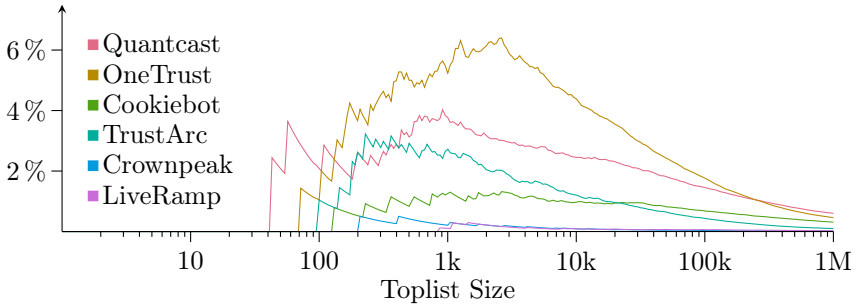


Figure 5.16: Cumulative CMP marketshare as a function of the toplist size (May 2020). This is a repetition of Figure 5.5 included for better comparison.

6

MEASURING THE IMPACT OF PRIVACY PREFERENCE SIGNALS

AUTHORS

Maximilian Hils, University of Innsbruck

CONFERENCE

Work in progress, to be submitted.

ABSTRACT

Since the passage of the General Data Protection Regulation (GDPR) in Europe, many websites employ cookie dialogs to obtain consent from users. Previous research has shown that AdTech regularly uses dark patterns in these dialogs to trick users into consenting. This paper goes beyond the user interface and sets out to analyze whether clicking "Reject All" in a cookie dialog does actually stop the data processing.

We perform manual and automated end-to-end measurements in which we first create personalized browser profiles, and then measure how different consent signals affect observed ad personalization. Our user study with 2093 website observations shows that many AdTech providers do indeed respect negative signals and stop showing personalized ads. We attempt to automate our measurements and instrument major browser engines from different vantage points using multiple crawling strategies. However, we find that the effects of privacy preference signals are hard to measure at scale due to AdTech's anti-bot measures.

While our main result is a positive one (AdTech respecting privacy preference signals), we suggest that this is simply because the risk of non-compliance currently outweighs the profit that could be gained from the small minority of users who do not give consent. With regulators enforcing easier opt-out mechanisms, measuring compliance will become increasingly necessary.

6.1 INTRODUCTION

For more than 20 years, stakeholders in the web ecosystem have tried to standardize signals that represent how users want their personal data to be processed. For example, the World Wide Web Consortium's Do Not Track (DNT) working group proposed a simple HTTP header users could send to opt out of tracking [1]. However, standardization failed after advertising companies withdrew from the standardization process and announced their intent to ignore the signal in 2013 [2]. More recently, in reaction to the European Union's General Data Protection Regulation (GDPR) enacted in 2016, a coalition of advertisers has successfully established a new standard, the Transparency and Consent Framework (TCF) [3]. Website owners use TCF to collect user consent in cookie dialogs, and then forward a standardized TCF signal to all embedded third parties. What all signals have in common is that they are *soft privacy* technologies. In contrast to *hard privacy* like encryption, they rely on the receiving entity to be trustworthy and respect the user's preferences.

A key question for privacy preference signals such as the TCF is whether third parties are compliant and do not start processing personal data unless they are given consent. After all, advertisers have monetary incentives to build extensive user profiles, which is at odds with respecting privacy-aware users' wishes. There is anecdotal evidence that compliance may not be taken too seriously by some players in the ecosystem: Pesch interviewed advertising companies and found that some only joined the TCF's vendor list because business partners required membership, claiming their own data processing would not require consent [4]. This suggests that data protection agencies should look for ways to detect illegal data processing. Privacy laws such as GDPR can only be effective if they are enforced by the respective authorities [5]. If compliance of AdTech vendors cannot be measured, there is little incentive to adhere to users' requests.

Researchers have tried to answer the question of compliance by checking whether the user's browser transmits the correct TCF signal in its HTTP requests to third parties [6, 7]. This method uncovers obvious privacy violations where the user's decision is already misrepresented by the (first party) website owner to third parties, but it does not help understand whether the signal-receiving third parties are compliant. Measuring this part of the ecosystem is difficult as server-to-server transfers are not visible to end users. In particular, the intransparent usage of TCF in real-time bidding (RTB) online ad auctions has been criticized by privacy advocates [8, 9]. In response to formal complaints, the Interactive Advertising Bureau, which governs the TCF standard, has launched a vendor compliance program in September 2021 [10, 11]. However, this approach has been dismissed by the complainants as being "unable to establish transparency and control" [12].

In this paper, we propose a method to monitor AdTech vendor compliance with end-to-end measurements of ad personalization (see Figure 6.1). Users first prime a browser profile with specific interests, for example by visiting

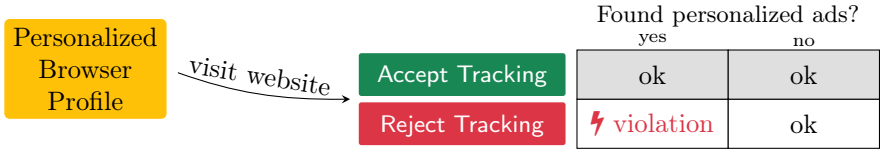


Figure 6.1: The high-level research approach in this paper is as follows: Using browser profiles primed with specific interests, a series of websites is visited. On each website, the user accepts or rejects all tracking. If no consent was given, the presence of personalized ads is taken as an indicator for unlawful data processing.

related websites and searching for relevant terms. This process collects cookies and other tracking identifiers, which can then be picked up by AdTech. In a second step, they visit general interest news websites where they either accept or reject all tracking in the consent dialog. If no consent is given to personalize ads and all vendors behave correctly, advertisements for the primed interests should be highly unlikely to appear. Seeing a statistically significant amount of personalized ads would demonstrate misconduct by the website or embedded AdTech parties.

The first contribution of this paper is a manual user study based on this method. Using 44 manual measurements of 50 news websites each, we show that rejecting a consent dialog stops most, but not all, ad personalization. Curiously, when users are instructed to also object to data processing based on legitimate interests (an advanced — usually well-hidden — opt-out mechanism available in TCF consent dialogs), a significant fraction of websites stops showing advertisements. This behavior is surprising insofar as the same websites were capable of showing non-personalized ads when the user did not perform this additional step.

The second contribution of this paper is the automation of the measurements. As part of our instrument validation, we show that the choice of Consent Management Provider (CMP) — the company that provides the standardized TCF cookie consent dialog for websites — influences the TCF signal that is communicated to third parties. We show that two major CMPs allow customers to hide the legitimate interest opt-out mechanism, a feature that is used by 10% and 15% of their customers. Comparing our automated and manual measurements, we find that our automated approach only captures a significantly smaller effect size, which can be attributed to measurement errors due to AdTech's anti-bot measures.

Second 6.2 provides background on relevant laws, the TCF, and related work. Second 6.3 describes our measurement methods. Section 6.4 presents the results. Section 6.5 discusses the results. Section 6.6 concludes the paper.

6.2 BACKGROUND

GDPR Since the GDPR [13] came into effect in 2018, all organizations offering goods or services to customers in the European Union must have a legal basis for processing their personal data. The GDPR defines possible

bases, most importantly *consent* and *legitimate interest* (Article 6.1). Consent is an opt-in mechanism. It must be "freely given, specific, informed and unambiguous indication of the data subject's wishes" (Recital 32) and "documented" (Article 7.1). Legitimate interest is an opt-out mechanism, but the firm must additionally demonstrate that the interest is legitimate, the processing is necessary, and balanced against the rights of data subjects (Article 6.1f). Recital 47 clarifies that individuals should reasonably expect the processing. One example for a legitimate interest would be fraud prevention, but Recital 47 adds that "processing of personal data for direct marketing purposes may be regarded as carried out for a legitimate interest."

TCF The TCF standard was developed by the Interactive Advertising Bureau Europe (IAB Europe), a coalition of advertisers, in reaction to the GDPR's data processing requirements. In short, it standardizes how users can consent to data processing. Under the TCF, website owners embed a Consent Management Provider (CMP) on their website, which displays a cookie consent dialog to visitors. The CMP records the user's decision and stores it as a standardized consent string. All other third parties embedded on the website can then use a standardized JavaScript API (`__tcfapi` [14]) to obtain a copy of this signal, which documents that they have a legal basis for their processing. In addition, the CMP discloses to the user that firms may be processing data based on legitimate interest, and provides the user with an option to object to such processing. A website wishing to implement the TCF independently must become a CMP, otherwise they can out-source this to an existing CMP. In reality, a handful of CMPs dominate the market [6].

The dialogs provided by CMPs are generally designed with AdTech's interest in mind. Existing research has repeatedly documented the dark patterns employed in consent dialogs to trick users into consenting [15, 16, 17, 18]. Consent rates of leading CMPs are reported to be well above 90% [19].

BEHAVIORAL ADVERTISING Closest to our research is the 2012 work of Balebako et al. [20], which measures the effect of opt-out cookies, blocking tools, and DNT headers on advertising personalization. They find that opt-out cookies and blocking tools reduce personalization, but sending a DNT header does not. We adopt their idea of using an engagement-related interest. AdTech's disregard of DNT is replicated without citing in a 2015 study [21].

The seminal work of Guha et al. [22] lays out fundamental challenges in measuring online advertising systems, but does not discuss the significance of anti-bot measures (which may not have been as much of an issue in 2010). The authors propose robust metrics to quantify the change in text ads between different browser profiles. Wills and Tatar [23] examine how (manually) controlled browsing influences the ads shown to users and interests shown in Ad Preference Managers. Hannak et al. [24] develop a methodology for measuring personalization in search results and investigate the influence of factors such as user agent and measurement location. Englehardt et al. [25] discuss engineering

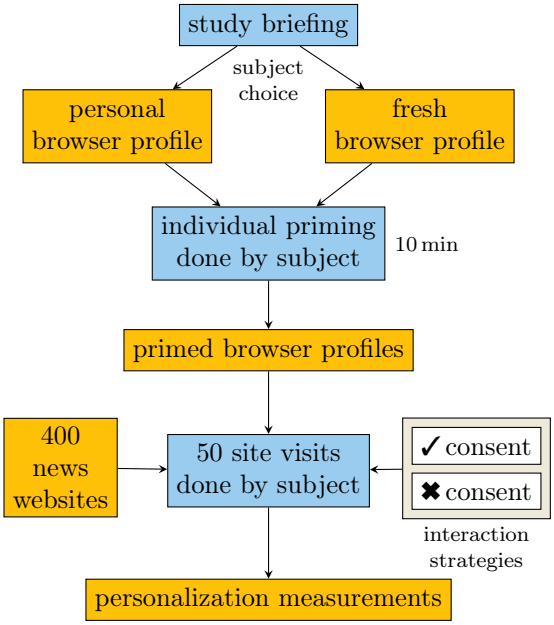


Figure 6.2: Manual Ad Personalization Measurements

challenges for web privacy measurements, which are particularly relevant in our context w.r.t. anti-bot measures. Barford et al. [26] develop a scalable crawler to capture 175k distinct display ads. Similar to this work, they create personalized browser profiles by visiting a set of interest-specific sites.

6.3 METHOD

We perform our personalization measurements both manually (where study participants control the browser) and in an automated fashion (where browser instrumentation is used). Section 6.3.1 describes the study design for the first approach, Section 6.3.2 describes the technical implementation of the second.

6.3.1 *Manual Measurements*

To test whether our measurement approach captures the expected effects, we performed manual measurements with the help of 29 graduate students at the University of Innsbruck, Austria in November 2021. Our study design is shown in Figure 6.2. Students first attended a two-hour lecture covering data protection law, GDPR terminology, and the TCF ecosystem as a regular part of a privacy and information security course. Next, we explained the objective of our study, the study design, and the potential side effects described below. The full study briefing document can be found in the appendix (Listing 6.1). We opted for such a detailed approach to make sure that students could make an informed decision on whether they would like to participate in the study. Students were then given the free choice to participate in the experiment in return for free beer (the majority of students opted not to go home).

Next, students were asked to make a choice on whether they would like to use their personal browser profile or whether they would like to use a separate/incognito profile for the study. Based on previous experiences we explained that using an existing profile could be helpful not to trigger anti-bot measures, but they were free to pick either option. Irrespective of their profile choice, students were instructed to disable any adblocking or anti-tracking browser extensions as well as VPNs.

To prime their browser profiles with a common interest, students browsed the web for ten minutes pretending to be someone who is planning to propose to their significant other. We picked this specific scenario for three reasons. First, only a small subset of the population is actively planning their engagement, which means that relevant ads are unlikely to be shown if nothing is known about the user. In contrast, ads for food or cars target much wider demographics and would be harder to attribute to personalization. Second, consumers looking for engagement rings have a high willingness to pay, which makes them an attractive audience to target. Finally, the IAB's Content Taxonomy (which is often used to target ads) has dedicated categories for this life event [27], which makes it easy for advertisers to reach out to this audience.

After priming their browser profiles, students were instructed to visit 50 news websites individually sampled for each student from the agof's Nov. 2021 toplist of 377 German digital media companies [28]. We picked this list instead of the more common Tranco [29] and Alexa [30] toplists because all included publishers serve ads, generally cater to wide audiences, and are active in the DACH (Germany (D), Austria (A), Switzerland (CH)) region. Each student was then assigned an experimental treatment in the form of either (1) unconditionally consenting to all data processing or (2) rejecting as much tracking as possible, i.e., also object to legitimate interests if such an option is provided by the website. We considered having students alternate between consenting and objecting, but deemed a constant strategy to produce more reliable measurements. For each visited page, students reported whether they observed personalized ads relating to the primed interest, generic ads, or no ads at all. All personalized ads were additionally documented with a screenshot.

Already during the measurements, we discovered that students in the reject group were seeing significantly fewer ads than expected, even though all adblockers were turned off. One hypothesis for this change was that the concentrated activity from the same IP range had been flagged by bot detection systems. We also considered that the lack of ads may have been triggered by the explicit objection to legitimate interests, which we only added after pretesting the study design.

To test both hypotheses, students were asked to repeat the measurement from home. We extended the existing study design with a third treatment in which users rejected tracking, but did not object to legitimate interests. Again, participation in this part was voluntary. Results were submitted anonymously

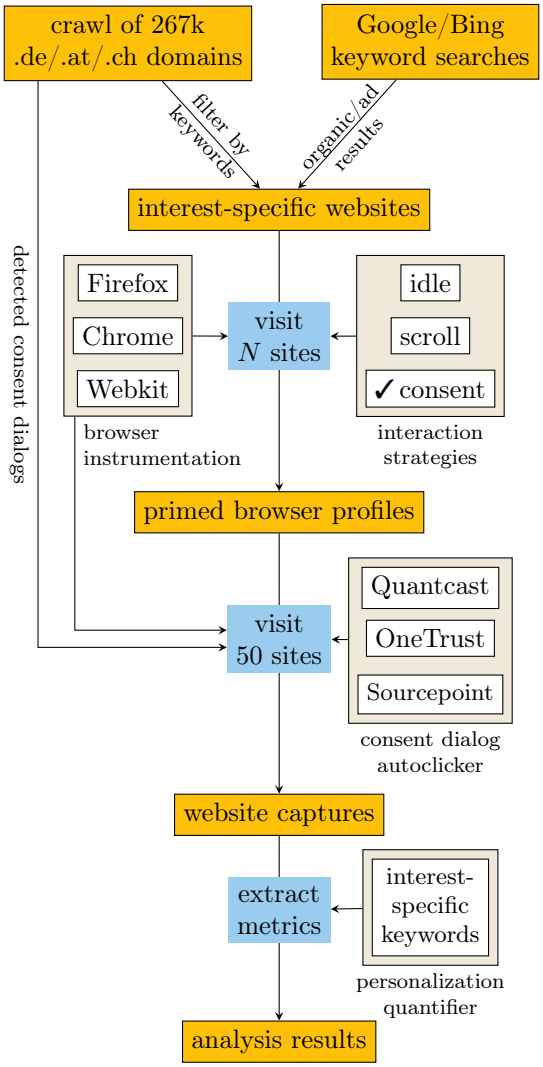


Figure 6.3: Automated End-to-End Ad Personalization Measurement Pipeline

so that students could be sure not to face negative consequences from not participating.

6.3.2 *Automated Measurements*

A distinct problem with the manual measurements is that attribution is difficult. While we can observe personalized ads, we cannot tell from a screenshot which third party is ultimately responsible for serving them. In contrast, an automated measurement allows us to record all traffic and investigate retrospectively. Additionally, users may receive seemingly personalized ads by chance, and we need a statistically significant amount of observations to confirm misbehavior. This motivates the development of an automated measurement platform. We provide a high-level overview in Figure 6.3 and explain the individual components shown there next.

GENERATING PRIMED BROWSER PROFILES. Before we can perform a measurement, we first need a primed browser profile with cookies and other tracking identifiers that indicates the topic we pretend to be interested in. We cannot reuse the same browser profile because the initial priming effect fades out as other (generic) websites are visited during the measurement. This means that we first need a way to automatically prime fresh browser profiles. To accomplish this, we visited multiple websites relating to our specific interest and automatically interacted with the pages to simulate human behavior. We implemented multiple basic interaction strategies such as mouse movement, scrolling, and clicking on internal links. For ethical reasons we wanted to avoid clicking on external links as those could be advertisements triggering payment.

SELECTING INTERESTS. We used engagement planning as the theme for all manual measurements to simplify the study instructions. For the automated measurements, we selected three additional interests from the IAB's content taxonomy that browser profiles could be primed for. First, we selected *weight loss* as a sensitive topic that pertains to the user's medical history. Our hypothesis here is that medical websites could be more likely to accept users' privacy preferences. Second, we chose *vaping* as this interest is explicitly listed as a sensitive topic in the IAB's taxonomy (weight loss is only listed as "Special Category Data"). Finally, we picked *SUVs* (sport utility vehicles) as a fourth — more innocuous — interest that appeals to a wider audience.

DETERMINING INTEREST-SPECIFIC WEBSITES. We next determined relevant websites for the chosen interests that could be visited for priming. We utilized both large-scale web crawls as well as search engine results to pick candidates.

To get a comprehensive picture of websites in the DACH area, we used the Chrome User Experience Report dataset to obtain a list of 267k .de/.at/.ch domains [31]. We then crawled the front page for each domain with a headless Chromium instance and stored all HTTP request headers and the final page text contents. Next, we systematically asked colleagues which terms they would associate with our four interests and compiled lists of on average 20 relevant keywords per interest. We additionally include relevant IDs from the IAB's content taxonomy. All page contents were then scanned using Hyperscan [32] and we used human judgment to manually determine a threshold that would qualify domains for being relevant to the given interest. A manual review of selected websites revealed a false-positive ratio of less than 10%, i.e., the vast majority of selected websites had a significant association with the respective interest. However, one downside of this approach is that website popularity is not taken into account, and websites in the long tail are overrepresented. For example, of the 299 .de/.at/.ch domains associated with the engagement interest, only 9 appear in the Tranco 1M toplist[1].

[1] The list we used is available at https://tranco-list.eu/list/6P7X.

We complemented our web crawls with targeted search engine queries using search terms proposed by colleagues during the keyword compilation. All queries were executed on Google and Bing, which are the two leading search engines in the DACH region (92% and 5% market share respectively). We recorded the first 100 organic results for each query. We did not click on any search advertisements for ethical reasons but added the respective websites to our list.

INSTRUMENTING BROWSERS To perform the browser profile priming and the measurements, we built a custom browser runner utilizing Microsoft's Playwright browser automation framework [33]. We opted for this approach as it allows us to instrument Google Chrome, Mozilla Firefox, Microsoft Edge, and Safari browser instances, whereas OpenWPM [34] only instruments Firefox. We did not use any browser's headless mode. To hide our instrumentation from anti-bot measures, we implemented basic cloaking mechanisms based on Jonker et al. [35]'s analysis of fingerprinting techniques. We manually verified that our modified browsers can pass openly available bot detection tools such as Google's ReCaptcha 3 when executed locally. However, we have no reliable means to test if we are picked up by AdTech's proprietary anti-bot mechanisms.

MEASUREMENT PLANS For each measurement, we defined a custom plan which states its key attributes (see Listing 6.2 in the appendix for a full example):

1. The location where measurements are done from.
2. The browser type and configuration (e.g., screen size).
3. The list of URLs and the page interaction strategy used for priming.
4. The list of URLs and the consent dialog interaction strategy for measurements.

We then executed browser runners that performed all browser instrumentation based on the assigned plan and uploaded raw results (DOM contents for all frames, cookies, screenshots, ...) into S3 object storage.

MEASUREMENT REGIONS We performed measurements from three different locations: Amazon Web Service's Frankfurt region, a cluster of virtual machines at the University of Innsbruck, and physical laptops at the University of Innsbruck. This is motivated as follows. Initially we intended to exclusively use a cloud computing platform as this would have allowed us to obtain a fresh IP address for each measurement. We considered this aspect to be important not to get any cross-contamination between measurements. However, our initial experiments yielded unsatisfactory results and we suspected that our IP address ranges may be blacklisted by AdTech companies. This led us to

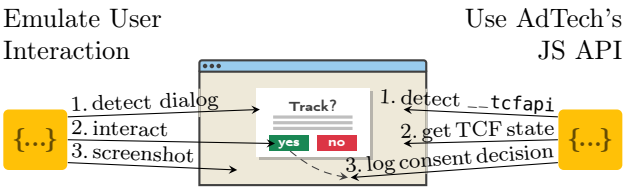


Figure 6.4: We measure the accuracy of our consent dialog automation by instrumenting websites on two different layers: First, we employ dialog-specific scripts to spoof user interaction. Second, we simultaneously interact with the website's TCF API to get notified of the user's decision from AdTech's perspective. This allows us to see if we instrumented the dialog correctly. We confirm inconsistent results with human raters.

transition to local virtual machines which share the same IP address space as students. As measurements still yielded unsatisfactory results, we suspected advanced fingerprinting of the underlying Linux VMs and transitioned to physical consumer laptops running Windows 10.

CONSENT DIALOG AUTOMATION On all pages we visited during the measurement phase we either accepted or rejected the consent dialog. While this task sounds straightforward, previous research had already demonstrated that it is surprisingly complex to solve even for the much simpler case of accepting dialogs [36]. In a measurement of 1426 websites, Matte et al. [6] automated all steps but the dialog interaction, which was performed by human operators to ensure its accuracy. As we set out to measure from cloud virtual machines, this was not an option for us. We attempted to adapt the existing Consent-O-Matic browser extension [37] for our use case but found its rule syntax to be too inflexible for objecting to legitimate interests in more complex dialogs. Instead, we re-crawled our list of 267k .de/.at/.ch domains to measure the popularity of specific CMPs and implemented custom dialog automation scripts for the three most popular CMPs (Quantcast, Sourcepoint, and OneTrust). To improve the accuracy of our method, we visited 500 websites per CMP on which we executed our dialog UI automation, and then simultaneously used AdTech's TCF JavaScript API (`__tcfapi`) to observe which decision is being communicated to third parties (see Fig 6.4). In other words, we instrumented each CMP from both the user and the AdTech side to detect mismatches between intended consent action and result. These observations could then be used to improve our automation and handle additional dialog configurations. We needed to repeat this process multiple times until all mismatches were confirmed to be a bug in the individual website itself and not in our instrumentation. In other words, a manual dialog interaction produced the same (wrong) TCF consent string on these websites.

MEASURING PERSONALIZATION Finally, we need automated means to detect the presence of personalized ads on each page. To do this at scale we used the list of keywords for each interest and counted their frequency on three different parts of the raw results: First, the concatenated DOM contents of

Date	Measurement	Fig.	# websites	CMPs
19.10.21	Dialog Automation Tests	6.6	3 000	3
19.10.–02.11.21	Cloud Measurements	6.7	36 000	3
09.11.–18.11.21	User Study	6.5	2 093	all
Nov. 21–Feb. 22	Local Measurements	6.7–6.8	72 000	3

Table 6.1: Overview of measurements made for this paper.

all frames on the page. Second, all HTTP requests and responses. Third, the browser's accessibility tree, which provides assistive technologies with textual descriptions of the web page.[2] We picked this differentiated approach to be able to capture all indicators (on the HTTP level), but also to filter out noise. For example, "Auto" — the German word for car — is not a useful indicator on the HTTP level (it appears as a property in most CSS stylesheets), but its presence in the accessibility tree is indicative of car ads. We then compute a simple personalization metric for each page by counting the number of matching keywords (each keyword is counted at most once per page). While a more sophisticated approach — one could envision training a machine learning model on screenshots in future work — would improve our results, the benefit of the current metric is that it is very easy to reason about.

6.3.3 *Research Ethics*

In fulfillment of approved ethical standards, we clearly communicated that participation in our manual measurement study was voluntary and anonymous. Students were extensively briefed on the general tracking ecosystem as well as the potential implications of participating in the study (see Listing 6.1). Participants could withdraw from the study at any point in time.

A broader concern for both manual and automated measurements is that we visit websites that include ads, which advertisers need to pay for. This problem is not unique to our method but an inherent property of web measurements. To limit the impact of our measurements, we did not interact with any of the displayed advertisements and instructed students not to do so either. We spaced out our measurements so as to not overload any servers. Considering that we only crawled each website a handful of times, we believe that we do not have significantly altered the economic ecosystem of the crawled websites.

6.4 RESULTS

We first describe the results from our manual measurement study, followed by the results from our automated measurements. Table 6.1 explains which measurements are used for the figures.

[2] The output format for the accessibility tree is browser-specific, which needs careful consideration when comparing results across browsers.

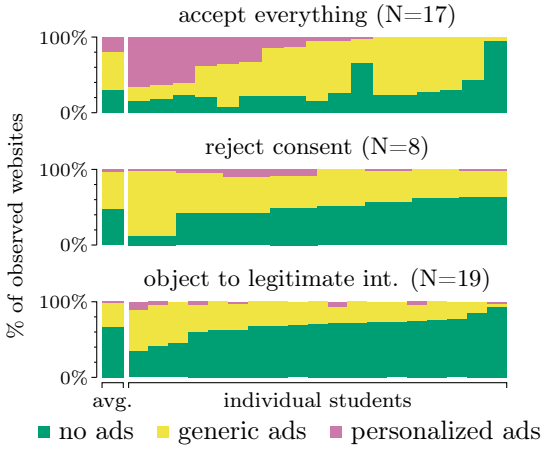


Figure 6.5: Manual measurements made by students reveal a high intra-group variance: While some students observed personalized ads on about 60% of the 50 websites they visited, about a third of students did not observe any personalized ads, even though all of them accepted all tracking.

6.4.1 *Manual Measurements*

In total, we received 17 measurements from students in class and 27 measurements from students at home. The increase in home measurements can be explained by the fact that many students had commitments after the initial lecture, but were still interested in seeing the effects of the priming personally. Each measurement consisted of up to 50 websites. After removing incomplete records, we record 2093 individual observations (Ø 47.6 websites/measurement).

The majority of students were Windows users (68%), followed by macOS (18%), and Linux (14%). 55% used Google Chrome, 18% Mozilla Firefox, 16% Microsoft Edge, 9% Safari, and 2% Chromium. More than two-thirds of students (73%) volunteered to use their existing browser profile, with many being curious about how pervasive the tracking would be. Others created a dedicated profile (16%) or used incognito mode (11%). Most participants performed the priming in German (86%), all others used English search terms.

Our manual measurements revealed two key results. **First, sending negative privacy preference signals significantly reduces the number of personalized ads that are seen.** More concretely, we observe personalized ads on on average 19.4% of websites when consenting to all data processing (see Figure 6.5). This number drops to 3.5% when not providing consent and 1.7% when objecting to legitimate interests. A Mann-Whitney U rank test confirms that the difference between accepting and simple rejecting is statistically significant, $U(N_{\text{accept}} = 17, N_{\text{reject}} = 8) = 230, p < 0.05$. Users who do the extra work to object to legitimate interests see significantly fewer ads (32% of websites) than those who only reject dialogs (52%), $U(N_{\text{reject}} = 8, N_{\text{legint}} = 19) = 22, p < 0.01$.

Second, we observe a high intra-group variance when measuring personalization. While some students that accepted tracking observed

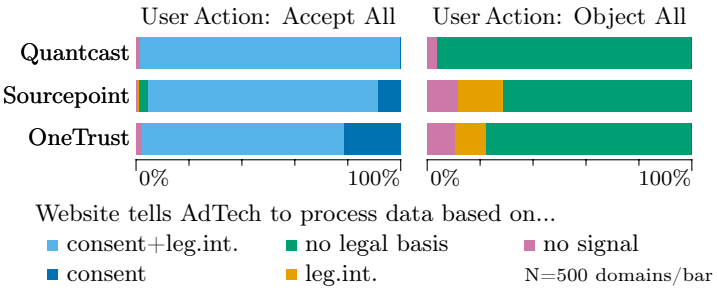


Figure 6.6: User action in the consent dialog vs. consent decision sent to AdTech for the three leading consent dialog providers.

personalized ads on as much as 60% of webpages they visited, a third of the students did not observe any personalization at all. This may be an indicator that the observed results highly depend on the quality of the priming process or that anti-bot measures play a significant role. Looking at the measurements where students objected to legitimate interests, we find that some students observed ads on more than 50% of the pages, but others saw ads on only 20% of pages. The refusal to serve ads supports the hypothesis that anti-bot measures may play a significant role. A chi-squared test confirms that the intra-group difference in not observing ads is statistically significant for students who objected to legitimate interests, $\chi^2(18, N = 941) = 83.53, p < 0.01$. Speaking to our hypothesis in Section 6.3.1 that concentrated activity from the same IP range may have distorted results, we find no statistically significant evidence that the measurement location had an effect on any treatment.

6.4.2 *Automated Measurements*

We first present results relating to the behavior of consent dialogs and then discuss the observed personalization from our automated measurements.

CONSENT DIALOG SETTINGS When testing our consent dialog automation, a surprising result for us was that (from their looks) identical consent dialogs would communicate different consent decisions to AdTech partners (see Figure 6.6). To make sure that these deviations are not an artifact of our automation, we manually confirmed all cases. In direct violation of the TCF's policies, we found that 16.7% of websites using Sourcepoint's dialog and 11.5% of websites using OneTrust's dialog do not provide European users with any means to object to legitimate interests. For 11.9% of Sourcepoint's and 10.8% of OneTrust's consent dialogs, we did not receive a consent signal via AdTech's JS API when trying to reject tracking. This was either because the dialog did not provide a reject option at all, or in a few cases because the website did not notify third parties that the user objected to legitimate interests[3].

[3] This behavior was likely well-intentioned. Some websites independently set an additional opt-out cookie to not include the CMP's JavaScript on subsequent page loads. The problem with this approach is that objection to legitimate interests becomes impossible.

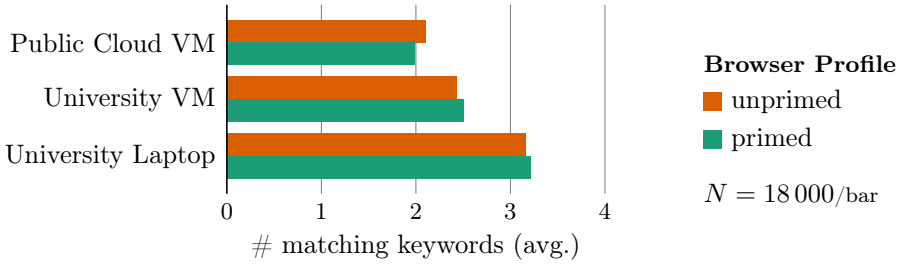


Figure 6.7: Observed ad personalization metrics for measurements made by different setups. The red bars represent the unpersonalized baseline, i.e., the average number of matching keywords in HTTP traffic if the browser profile has not been primed. In all cases, the actual personalization effect — the difference between the red and green bars — is small.

We also found a surprising amount of Sourcepoint dialogs (3.6%) which were misconfigured to the extent that clicking "Accept all" (or a variation thereof) resulted in a negative consent signal. We can also report encouraging results in that 21.3% of OneTrust customers and 8.5% of Sourcepoint customers actively prohibit third parties from using legitimate interest as a legal base for processing, even if the user accepts all tracking.

OBSERVED PERSONALIZATION While our manual measurements show strong personalization effects after priming, the same unfortunately cannot be said about our automated measurements. Recall that our personalization metric counts the occurrence of keywords associated with the primed topics. The red bars in Figures 6.7 and 6.8 represent the non-personalized baseline (keywords matching by chance), and the difference between grouped red and green bars describes the personalization effect (we give full consent in both cases). Looking at Figure 6.7, we see a *negative* personalization effect on cloud-hosted virtual machines (this effect is statistically significant). In other words, if we first visit websites related to engagements, we are *less* likely to observe related keywords on subsequent generic page visits.

Why is that so? Our best explanation for the decrease in matching keywords is that the priming process triggers anti-bot measures, which then affect the measured sites. The effect is consistent across time of day, day of measurement, and browsers. Each measurement was done on a new virtual machine with a new IP address so that we can exclude cross-contamination between measurements as a contributing factor. The decrease in keywords is also consistent across primed interests and priming strategies. This makes us assume that it's primarily based on a combination of IP address reputation and fingerprinting techniques.

Switching to local virtual machines with constant University IP addresses, we find that keywords are 21% more likely to appear, but we also see a similar increase for the unprimed case. This may be because some advertisers personalize based on IP addresses and not based on cookies, or because the IP address range we are crawling from is generally seen as more trustworthy, which results in more ads being served. Using physical laptops yields another 27%

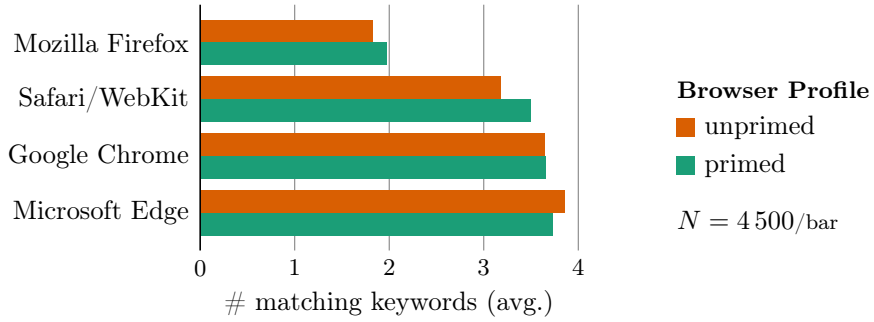


Figure 6.8: Observed ad personalization metrics for measurements made by different browsers from the "University Laptop" setup.

increase, which strengthens the hypothesis that anti-bot measures significantly influence results. However, we still observe almost no effect from priming on personalization.

Turning to Figure 6.8, which only considers the measurements on physical laptops, we find that the choice of browser also significantly impacts measurements from the same devices. Most strikingly, we observe 50% fewer keywords in Mozilla Firefox. Again, the impact of browser choice seems to be much more significant than the impact of our actual priming process. We discuss implications in the next section.

6.5 DISCUSSION

Section 6.5.1 discusses our measurement methodology. Section 6.5.2 discusses implications for privacy policy.

6.5.1 *Methodological Implications*

CONSENT DIALOG AUTOMATION As more websites delay loading third parties until the user made a consent decision, privacy web measurements need to implement consent dialog instrumentation to observe all tracking [36]. This task can be done manually with human operators [6], which covers arbitrary websites and ensures high accuracy, but does not scale well. Alternatively, Jha et al. [36] proposes the use of generic heuristics to accept dialogs, trading off accuracy for scalability. Finally, custom scripts developed for specific dialog implementations can provide both scale and accuracy, but do not cover all websites. We think this approach is most suitable for measurements like ours. By focusing on the three most popular CMPs, we automate 53% of DACH websites implementing TCF with much higher reliability compared to existing work. For example, Jha et al. [36] report that Consent-O-Matic achieves a 24% dialog acceptance rate for 100 randomly picked German websites. Their Priv-Accept extension is successful on more than 50% of websites but only supports accepting dialogs. In contrast, we were able to iteratively tweak our instrumentation based on the feedback we received via TCF's AdTech API

(see Figure 6.4), which yields a 98.5% success rate for accepting dialogs of the specific CMPs (see Figure 6.6).

We note that all automation approaches require continuous adjustments and revalidation as dialogs change over time. Hils et al. [38] observed 38 changes to only Quantcast's dialog over the span of three years.

BROWSER CHOICE We find that the choice of browser has a significant influence on the observed personalization metrics, which is in line with existing research [39]. The significant decrease in keywords for our Firefox instances indicates that they are easily detected by anti-bot measures despite our attempts at concealing them. We do not know the exact root cause, but it may not be a coincidence that this affects the browser that is very commonly used for instrumentation. For example, OpenWPM uses Firefox exclusively [34]. In contrast, Safari/WebKit is rarely instrumented, but it is the only browser where we observe a notable priming effect. This reaffirms the importance of using different browsers for measurements, at least when looking at ecosystems where anti-bot measures are popular. We simply wouldn't have noticed the missing keywords had we only instrumented Firefox.

MEASURING PRIMING EFFICACY In the first part of this study, we planned to measure how variables such as the data source for interest-specific websites (keyword search vs. crawl), the number of visited pages, or the page interaction strategy would influence the efficacy of the priming process. A key problem we encountered here is that (1) reusing our own IP address introduces cross-contamination, but (2) many advertisers block the IP address space of cloud providers. Future research will need to either accept cross-contamination for measurements (and not make strong statements on priming efficacy) or obtain access to a large number of reputable IP addresses.

MEASURING (EFFECTS ON) PERSONALIZATION While we planned to measure the effect of consent signals on personalization, we must concede that our automated measurements failed to capture the underlying personalization effect in the first place. We could of course easily report statistically significant results with a bit of cherry-picking, but the effect size simply does not compare to what we observed in our manual user study. As such, one needs to exercise caution in drawing conclusions from the data. AdTech's anti-bot measures provide no direct feedback by design, which makes it easy to miss vital factors such as the Firefox issues we described above. Performing manual measurements as a comparison baseline is a good practice in this context.

As part of this research, we learned that AdTech by nature is highly adversarial when it comes to web measurements. Ad fraud is a widespread phenomenon [40], and vendors have rightfully deployed extensive countermeasures that also affect researchers. Again, manual measurements offer a pragmatic recourse.

6.5.2 *Privacy Implications*

IMPACT OF PRIVACY PREFERENCE SIGNALS Even though enforcement of the GDPR is often criticized as lacking [5, 41], our manual measurements show that most websites respect users' choices nonetheless and stop showing personalized ads when no consent is given. This is an encouraging result, but it should be celebrated with caution. With current consent rates of above 90% (see Section 6.2), it may make little sense for AdTech to not respect preference signals, because the risk of non-compliance outweighs the profit that could be gained from the small minority of users who do not give consent. A similar argument was already brought up by Szoka [42] in discussions about DNT: AdTech may tolerate a small number of "free-riders" as long as it does not exceed their "maximum acceptable loss threshold." This hypothesis is further supported by our observation that many websites stop showing advertisements altogether when the user objects to legitimate interests, a right that we estimate is only exercised by a tiny fraction of users. Once regulators start to enforce the GDPR's mandate that consent dialogs need to make withdrawal as easy as giving consent [13, Art. 7], we may observe a shift in AdTech's behavior. It would be surprising to see AdTech liberally employ dark patterns in their consent dialogs, but then behave in an exemplary manner when it comes to respecting preferences.

EFFECTIVENESS OF SOFT PRIVACY Measuring the presence of personalized ads as done in this paper can only catch some forms of non-compliant behavior. All data processing that is done for purposes other than advertising remains out of view. It may be tempting to advocate for solutions from the domain of *hard privacy* instead, for example blocking third-party cookies by default. However, we already see advertisers circumventing this approach with first-party cookie syncing [43, 44]. It remains doubtful whether stricter measures could comprehensively prevent tracking, at least without breaking a significant fraction of the existing web. In this context, we see soft privacy technologies as a necessary complement. However, they require effective enforcement to not be ignored by others.

PRACTICAL ENFORCEMENT While we are happy that most personalized ads disappear with negative consent signals, some unwanted personalization remains. This brings up the question of what should be done about noncompliance. Unfortunately, a foundational problem here is attribution. While our manual study provided us with screenshots of each offending ad, it's likely that neither the website owner nor the company providing the ad contents is directly to blame for illegal targeting. s Determining which specific AdTech vendor is misbehaving requires manual detective work, for example by looking at the underlying HTTP traffic. This process is further aggravated by the fact that presumably personalized ads may have been a coincidence, and we need statistical evidence to not end up chasing ghosts. Our automated measurements

would have provided a basis for this direction but failed to capture the desired effects. This leaves this part of the paper with a meta result: Measuring the impact of privacy preference signals at scale is a very hard problem to solve. This is bad news for privacy advocates and data protection agencies, who need to rely on laborious manual methods to keep tabs on AdTech. A possible way forward would be to mandate transparency for targeting decisions. For example, AdTech could provide audit APIs to disclose why particular ads are shown to users.

CMPS VIOLATING TCF POLICY As part of our automated measurements, we found that Sourcepoint and OneTrust allow their customers to hide the opt-out button for legitimate interest in their consent dialogs. This option — used by more than 10% of each companies' customers — is a direct violation of TCF's own policies [45, Chapter II 5(4) and Appendix B]. This raises questions as to whether the IAB is following its pledge to "take reasonable steps to periodically review and verify a CMP's compliance" [45, Chapter II 9(1)].

6.6 CONCLUSION

We perform manual and automated measurements to test whether the privacy preferences stated in a consent dialog influence the observed personalization of ads on the web. We find that most websites stop showing personalized ads when no consent is given. However, the difficulties faced in automating our measurements show that compliance detection is hard to scale. This threatens the efficacy of the TCF privacy preference signal. If more users are enabled to make use of their rights, AdTech vendors will be increasingly tempted to ignore negative consent signals.

REFERENCES

[1] World Wide Web Consortium. Tracking Protection Working Group, 2011. https://www.w3.org/2011/tracking-protection/.

[2] Maximilian Hils, Daniel W. Woods, and Rainer Böhme. Privacy Preference Signals: Past, Present and Future. *Proceedings on Privacy Enhancing Technologies*, (4), 2021. https://doi.org/10.2478/popets-2021-0069.

[3] IAB Europe. What is the Transparency and Consent Framework (TCF)?, 2020. https://iabeurope.eu/transparency-consent-framework/.

[4] Paulina Jo Pesch. Drivers and Obstacles for the Adoption of Consent Management Solutions by Ad-Tech Providers. In *Euro S&P Workshop on Consent Management in Online Services (COnSeNT)*, 2021. https://doi.org/10.1109/EuroSPW54576.2021.00034.

[5] Irish Council for Civil Liberties. Europe's enforcement paralysis: ICCL's 2021 GDPR report, September 2021. https://www.iccl.ie/wp-content/uploads/2021/09/Europes-enforcement-paralysis-2021-ICCL-report-on-GDPR-enforcement.pdf.

[6] Célestin Matte, Nataliia Bielova, and Cristiana Santos. Do Cookie Banners Respect my Choice? Measuring Legal Compliance of Banners from IAB Europe's Transparency and Consent Framework. In *IEEE Symposium on Security and Privacy*, pages 791–809. IEEE, 2020. https://doi.org/10.1109/SP40000.2020.00076.

[7] Koen Aerts. Cookie Dialogs and Their Compliance. Master's thesis, Open University of the Netherlands, July 2021. https://www.open.ou.nl/hjo/supervision/2021-koen-aerts-msc-thesis.pdf.

[8] Brave Browser. Updates & timeline for Brave's work to fix "RTB" adtech, January 2020. https://brave.com/rtb-updates/.

[9] Irish Council for Civil Liberties. Data Protection Authority investigation finds that the IAB Transparency and Consent Framework infringes the GDPR, October 2020. https://www.iccl.ie/news/apd-iab-findings/.

[10] IAB Europe. IAB Europe Launches New TCF Vendor Compliance Programme, August 2021. https://iabeurope.eu/all-news/update-on-the-belgian-data-protection-authoritys-investigation-of-iab-europe/.

[11] IAB Europe. Update On The Belgian Data Protection Authority's Investigation Of IAB Europe, November 2021. https://iabeurope.eu/blog/iab-europe-launches-new-tcf-vendor-compliance-programme/.

[12] Irish Council for Civil Liberties. IAB Europe can't audit what 1000+ companies that use its TCF system do with our personal data, January 2022.

https://www.iccl.ie/digital-data/iab-europe-cant-audit-what-1000-companies-that-use-its-tcf-system-do-with-our-personal-data/.

[13] European Union. Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation). *Official Journal L110*, 59:1–88, May 2016.

[14] IAB Europe. Consent Management Platform API, 2021. https://github.com/InteractiveAdvertisingBureau/GDPR-Transparency-and-Consent-Framework/blob/841f7ef/TCFv2/IAB%20Tech%20Lab%20-%20CMP%20API%20v2.md.

[15] Arvind Narayanan, Arunesh Mathur, Marshini Chetty, and Mihir Kshirsagar. Dark Patterns: Past, Present, and Future. *ACM Queue*, 18(2): 67–92, April 2020. https://doi.org/10.1145/3400899.3400901.

[16] Midas Nouwens, Ilaria Liccardi, Michael Veale, David Karger, and Lalana Kagal. Dark Patterns after the GDPR: Scraping Consent Pop-Ups and Demonstrating Their Influence. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. ACM, 2020. ISBN 9781450367080. https://doi.org/10.1145/3313831.3376321.

[17] Arunesh Mathur, Gunes Acar, Michael J. Friedman, Eli Lucherini, Jonathan Mayer, Marshini Chetty, and Arvind Narayanan. Dark Patterns at Scale. *Proceedings of the ACM on Human-Computer Interaction*, 3 (CSCW):1–32, November 2019. https://doi.org/10.1145/3359183.

[18] Christoph Bösch, Benjamin Erb, Frank Kargl, Henning Kopp, and Stefan Pfattheicher. Tales from the Dark Side: Privacy Dark Strategies and Privacy Dark Patterns. *Proceedings on Privacy Enhancing Technologies*, 2016(4):237–254, 2016. https://doi.org/10.1515/popets-2016-0038.

[19] NOYB – European Center for Digital Rights. noyb aims to end "cookie banner terror" and issues more than 500 GDPR complaints, 2021. https://noyb.eu/en/noyb-aims-end-cookie-banner-terror-and-issues-more-500-gdpr-complaints.

[20] Rebecca Balebako, Pedro G. Leon, Richard Shay, Blase Ur, Yang Wang, and Lorrie Faith Cranor. Measuring the Effectiveness of Privacy Tools for Limiting Behavioral Advertising. In *In Web 2.0 Workshop on Security and Privacy*, 2012. https://lorrie.cranor.org/pubs/EffectivenessBA.pdf.

[21] Juan Miguel Carrascosa, Jakub Mikians, Ruben Cuevas, Vijay Erramilli, and Nikolaos Laoutaris. I Always Feel Like Somebody's Watching Me. Measuring Online Behavioural Advertising. In *Proceedings of the 11th*

ACM Conference on Emerging Networking Experiments and Technologies, CoNEXT '15. ACM, 2015. https://doi.org/10.1145/2716281.2836098.

[22] Saikat Guha, Bin Cheng, and Paul Francis. Challenges in Measuring Online Advertising Systems. In *Proceedings of the Internet Measurement Conference 2010*, IMC '10. ACM, 2010. https://doi.org/10.1145/1879141.1879152.

[23] Craig E. Wills and Can Tatar. Understanding what they do with what they know. In *Proceedings of the 2012 ACM workshop on Privacy in the electronic society*, WPES '12. ACM, 2012. https://doi.org/10.1145/2381966.2381969.

[24] Aniko Hannak, Piotr Sapiezynski, Arash Molavi Kakhki, Balachander Krishnamurthy, David Lazer, Alan Mislove, and Christo Wilson. Measuring Personalization of Web Search. In *Proceedings of the 22nd International World Wide Web Conference*, WWW '13, Rio de Janeiro, Brazil, May 2013. https://doi.org/10.1145/2488388.2488435.

[25] Steven Englehardt, Christian Eubank, Peter Zimmerman, Dillon Reisman, and Arvind Narayanan. Web Privacy Measurement: Scientific principles, engineering platform, and new results, 2014. https://www.cs.princeton.edu/~arvindn/publications/WebPrivacyMeasurement.pdf.

[26] Paul Barford, Igor Canadi, Darja Krushevskaja, Qiang Ma, and S. Muthukrishnan. Adscape: Harvesting and Analyzing Online Display Ads. In *Proceedings of the 23rd International Conference on World Wide Web*, WWW '14. ACM, 2014. https://doi.org/10.1145/2566486.2567992.

[27] IAB Tech Lab. Interactive Advertising Bureau Content Taxonomy, 2021. https://iabtechlab.com/standards/content-taxonomy/. Version 3.0.

[28] Arbeitsgemeinschaft Onlineforschung e.V. Monthly reports of daily digital facts, November 2021. https://www.agof.de/en/studien/daily-digital-facts/monatsberichte/.

[29] Victor Le Pochat, Tom van Goethem, Samaneh Tajalizadehkhoob, Maciej Korczynski, and Wouter Joosen. Tranco: A Research-Oriented Top Sites Ranking Hardened Against Manipulation. In *26th Annual Network and Distributed System Security Symposium*, NDSS '19. The Internet Society, 2019. https://doi.org/10.14722/ndss.2019.23386.

[30] Alexa. Top 1M sites toplist, 2022. https://www.alexa.com/topsites.

[31] Google. Chrome User Experience Report, November 2021. https://developers.google.com/web/tools/chrome-user-experience-report. Ver. 202101.

[32] Xiang Wang, Yang Hong, Harry Chang, KyoungSoo Park, Geoff Langdale, Jiayu Hu, and Heqing Zhu. Hyperscan: A Fast Multi-pattern Regex Matcher for Modern CPUs. In *16th USENIX Symposium on Networked Systems Design and Implementation*, NSDI '19, pages 631–648, Boston, MA, 2019. USENIX. ISBN 978-1-931971-49-2. https://www.usenix.org/conference/nsdi19/presentation/wang-xiang.

[33] Microsoft. Playwright: A Framework for Web Testing and Automation, 2022. https://playwright.dev. Ver. 1.17.0.

[34] Steven Englehardt and Arvind Narayanan. Online Tracking: A 1-Million-Site Measurement and Analysis. In *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security*, CCS '16, pages 1388–1401. ACM, 2016. ISBN 9781450341394. https://doi.org/10.1145/2976749.2978313.

[35] Hugo Jonker, Benjamin Krumnow, and Gabry Vlot. Fingerprint Surface-Based Detection of Web Bot Detectors. In *Lecture Notes in Computer Science*, pages 586–605. Springer International Publishing, 2019. https://doi.org/10.1007/978-3-030-29962-0_28.

[36] Nikhil Jha, Martino Trevisan, Luca Vassio, and Marco Mellia. The Internet with Privacy Policies: Measuring The Web Upon Consent. September 2021. https://arxiv.org/abs/2109.00395v1.

[37] Clemens Nylandsted Klokmose, Janus Bager Kristensen, and Rolf Bagge. Consent-O-Matic, May 2021. https://github.com/cavi-au/Consent-O-Matic. Ver. 0.9.4.

[38] Maximilian Hils, Daniel W Woods, and Rainer Böhme. Measuring the Emergence of Consent Management on the Web. In *Proceedings of the Internet Measurement Conference 2020*, IMC '20. ACM, 2020. https://doi.org/10.1145/3419394.3423647.

[39] Syed Suleman Ahmad, Muhammad Daniyal Dar, Muhammad Fareed Zaffar, Narseo Vallina-Rodriguez, and Rishab Nithyanand. Apophanies or Epiphanies? How Crawlers Impact Our Understanding of the Web. In *Proceedings of The Web Conference 2020*. ACM, apr 2020. https://doi.org/10.1145/3366423.3380113.

[40] Brett Stone-Gross, Ryan Stevens, Apostolis Zarras, Richard Kemmerer, Chris Kruegel, and Giovanni Vigna. Understanding Fraudulent Activities in Online Ad Exchanges. In *Proceedings of the Internet Measurement Conference 2011*, IMC '11. ACM, 2011. https://doi.org/10.1145/2068816.2068843.

[41] Josephine Wolff and Nicole Atallah. Early GDPR Penalties: Analysis of Implementation and Fines Through May 2020. *Journal of Information Policy*, 11(1):63–103, jan 2021. https://doi.org/10.5325/jinfopoli.11.2021.0063.

[42] Berin Michael Szoka. The Paradox of Privacy Empowerment: The Unintended Consequences of 'Do Not Track'. *SSRN Electronic Journal*, 2013. https://doi.org/10.2139/ssrn.2318146.

[43] Quan Chen, Panagiotis Ilia, Michalis Polychronakis, and Alexandros Kapravelos. Cookie Swap Party: Abusing First-Party Cookies for Web Tracking. In *Proceedings of the Web Conference 2021*. ACM, 2021. https://doi.org/10.1145/3442381.3449837.

[44] Nurullah Demir, Daniel Theis, Tobias Urban, and Norbert Pohlmann. Towards Understanding First-Party Cookie Tracking in the Field. 2022. https://arxiv.org/abs/2202.01498v2.

[45] IAB Europe. IAB Europe Transparency & Consent Framework Policies, 2021. https://iabeurope.eu/wp-content/uploads/2021/09/TransparencyConsentFramework-_Policies_version_TCF-v2.0-2021-06-22.3.4.docx.pdf.

APPENDIX

Listing 6.1: Study Briefing Document (2 pages)

Consent Management Experiment

MAXIMILIAN HILS, UIBK SECURITY AND PRIVACY LAB

If at any point you have any questions about the study, please reach out to maximilian.hils@uibk.ac.at.

1 STUDY OBJECTIVE

The objective of this study is to determine whether accepting or rejecting cookies has an influence on the advertisements that are shown on websites. We want to test how providing consent influences the amount of personalization seen in ads.

2 RESEARCH ETHICS

Participation in this study is voluntary. All data you submit will be published as supplementary material of the research paper documenting this study. You may withdraw from the study at any point in time and delete the data you have submitted so far.

As part of this study, you will be asked to visit websites that relate to engagement planning. Due to the pervasiveness of adtech tracking, your partner or other members in your household may see wedding/engagement-related advertisements on their devices in the next few days. If this is uncomfortable for you, do not participate in this study.

As part of this study, you will be asked to submit website screenshots. These screenshots contain advertisements, which may show contents related to your personal interests. You may skip any screenshot you do not wish to share.

3 STUDY DESIGN

This study consists of three parts. First, you will *setup* your browser to be ready for the measurements. Second, you will *prime* your browser profile cookies by visting websites that relate to engagement planning. Finally, you will visit websites randomly drawn from the most popular websites in Austria and *measure* if there are advertisements relating to this topic.

3.1 Setup

(1) **Decide which browser and browser profile** you want to use. To increase the accuracy of our study, we would appreciate if you could use your normal browser profile (which already carries cookies that are associated with human-like behavior), but you may also create a new browser profile if you are uncomfortable with using your normal profile.[1] Unless you are using Safari, please create a profile and do not use incognito mode. Incognito mode is commonly detected by advertising companies. You can delete the profile after completion.
(2) **Deactivate all adblocking and anti-tracking** browser extensions.
Disable your VPN if you use any.

[1] **Chrome/Edge:** Start Chrome and click on the user icon in the top right. **Firefox:** Enter `about:profiles` in the URL bar. **Chromium:** `chromium --user-data-dir=/tmp/chrome` **Safari:** Not possible, use incognito mode instead.

Author's address: Maximilian Hils, UIBK Security and Privacy Lab, maximilian.hils@uibk.ac.at.

3.2 Priming Phase

Start priming the browser profile **for ten minutes**:

(3) **Pretend to be someone who is planning to propose to their significant other**. Search for engagement rings, diamonds, proposal tips, wedding things, and so on to make advertisers believe that you are interested in this topic.
   - Do everything in German if possible, otherwise in English.
   - If you encounter any cookie prompts, accept them.
   - Make sure that you actively engage (scroll, click around, search on site, ...) with websites for more than a few seconds. This is a common method to detect bots.
   - You may also improve priming by including searches on Instagram/Facebook/etc.
   - Please do not click on advertisements.

3.3 Measurement Phase

We want to measure whether your priming has an impact on the ads you are seeing now.
Take the last digit of your matriculation number. If it is a...

- **0, 1, or 2**: You are lucky: Your strategy will be to *accept* all cookie dialogs in this phase.
- **3, 4, or 5**: Your strategy will be to *reject* all cookie dialogs in this phase, but not object to legitimate interests. This is typically done by clicking a "Reject All" button in the dialog (if one exists), or by clicking "More Options" and then something like "Save & Exit". Imagine a regular human being who does not want to be tracked, but also wants the dialog to go away.
- **6, 7, 8, or 9**: Your strategy will be to *reject all* cookie dialogs and object to all data processing based on legitimate interest ("Berechtigtes Interesse"). In other words, you want to exercise the maximum amount of privacy options available to you. This often means clicking "More Options", then searching for a legitimate interest button, and then manually objecting to each purpose.

For simplicity, you perform the same action (either accepting or rejecting cookies) on all webpages.
Next, do the following:

(4) **Open** https://uploads.hi.ls/2021-11/consent-exercise.html and select the correct sheet.
(5) **Take the first participant ID** where no measurements have been entered yet.
(6) **Enter your metadata** (browser, operating system, primary language) into the first row with your participant ID.
(7) **Visit each website** listed under your participant ID:
   (a) **Complete the consent dialog** if it exists and if your action is possible (some dialogs have no reject button – don't do anything then if you are in one of the reject groups).
   (b) **Check if there are advertisements,** and if they are personalized.
   (c) **Make a screenshot** of the browser window if (and only if) there are personalized ads, i.e., ads related to engagement/marriage things.[2] In the spreadsheet, at the very bottom, select the sheet belonging to your participant ID and upload the screenshot there.
   (d) **Enter your observations** into the spreadsheet.

After completing the measurement phase, make sure to reactivate your adblocking browser extensions. Thank you! :-)

[2] **Windows/Linux:** Focus the browser window, press `Alt+Print` (`Alt+Druck`). The screenshot is now in your clipboard, and can be pasted into the spreadsheet. **macOS:** Press `Shift`, `Command`, and `4`. In the spreadsheet menu, click `Insert`, then `Image`, and select the screenshot file on your desktop.

Listing 6.2: Measurement Plan Example

```
{
    "version": 2,
    "id": "test/paper-example",  // unique measurement ID
    "region": "aws/eu-central-1",  // runner location

    // browser configuration
    "device": {
        "type": "chromium",
        "options": {
            "headless": false,
            "channel": "chrome", // use official Google binaries
            "userAgent": "[...]",
            "viewport": {
                "width": 1536,
                "height": 763
            }
        }
    },
    "locale": "de-DE,de;q=0.9,en-GB;q=0.8,en;q=0.7",
    "timezone": "Europe/Berlin",
    "concurrency": 2, // number of parallel tabs

    // measurement strategy
    "prime": {  // visit engagement-related .de/.at/.ch URLs
        "urls": {
            "collection": "engagement_dach",
            "pages": 25
        },
        "strategy": "idle" // scroll, browse, ...
    },
    "measure": {  // visit domains with supported CMPs
        "urls": {
            "collection": "cmp_dach",
            "pages": 50
        },
        "strategy": "consent_reject"  // accept, idle, ...
    },

    // result storage (S3)
    "log": {
        "screenshot": "full", // including scrolling
        "contents": true, // final rendered DOM tree
        "cookies": true, // browser profile cookie jar
        "accessibility_tree": true,
        "har": true,  // HTTP Archive
        "console": true  // JavaScript console output
    }
}
```

7

CONFLICTING PRIVACY PREFERENCE SIGNALS IN THE WILD

AUTHORS

Maximilian Hils, University of Innsbruck
Daniel Woods, University of Innsbruck
Rainer Böhme, University of Innsbruck

CONFERENCE

15th International Conference on Computers, Privacy & Data Protection (CPDP)
23–25 May 2022, Brussels, Belgium.

ABSTRACT

Privacy preference signals allow users to express preferences over how their personal data is processed. These signals become important in determining privacy outcomes when they reference an enforceable legal basis, as is the case with recent signals such as the Global Privacy Control and the Transparency & Consent Framework. However, the coexistence of multiple privacy preference signals creates ambiguity as users may transmit more than one signal. This paper collects evidence about ambiguity flowing from the aforementioned two signals and the historic Do Not Track signal. We provide the first empirical evidence that ambiguous signals are sent by web users in the wild. We also show that preferences stored in the browser are reliable predictors of privacy preferences expressed in web dialogs. Finally, we provide the first evidence that popular cookie dialogs are blocked by the majority of users who adopted the Do Not Track and Global Privacy Control standards. These empirical results inform forthcoming legal debates about how to interpret privacy preference signals.

7.1 INTRODUCTION

Privacy laws like GDPR and CCPA empower users, at least in theory, to control how their personal data is processed. To do so, individuals must be able to communicate their privacy preferences with data controllers. A web standard for communicating privacy preference signals would make this convenient and easy. However, the literature suggests coordinating senders and recipients to adopt one standard has failed multiple times due to the competing interests of stakeholders [9]. The same competing interests lead stakeholders to propose different signals, leading to the coexistence of multiple signals. Users may transmit more than one signal and thereby express conflicting or ambiguous preferences, which creates uncertainty over which legal rules apply.

Multiple signals can be sent when signals are collected at different technical layers. In this study, we focus on the two dominant ways for users to express privacy preferences: on individual websites and globally in their browser. The first approach is chosen by the Transparency & Consent Framework (TCF), a standard developed by the Interactive Advertising Bureau and adopted by hundreds of ad-tech vendors and thousands of websites [15]. The second approach was chosen by the Do Not Track (DNT) mechanism [17] and also the Global Privacy Control (GPC), which now boasts over 40 million users [6].

The possibility of users sending multiple signals raises questions about legal interpretation under both the CCPA and the GDPR. Such questions can only be answered by legal analysis. Such scholarship would be supported by first establishing which signals users send in the wild, the problem addressed by this paper. We observe 16k impressions on websites that embed TCF dialogs and simultaneously detect the presence of a DNT and/or GPC signal.

Our results uncover a number of sources of ambiguity not previously identified in the literature. First, an industry standard dialog for collecting TCF signals is blocked by 27% of users, and this percentage rises to 50%/73% of users with DNT/GPC turned on. Second, users who send a GPC signal are two times more likely to withhold consent than other users, which suggests the signal captures genuine privacy preferences. Finally, even though they are more likely to not give consent, 73% of users with GPC turned on still consent to being tracked by clicking "I Accept" in a TCF consent dialog. This shows that conflicting signals are a reality.

Section 7.2 provides background on relevant laws and signals. Section 7.3 describes our research design. Section 7.4 presents the results. Section 7.5 discusses the results and suggests directions for future work. Section 7.6 concludes the paper.

7.2 BACKGROUND

We split the background into laws and privacy preference signals.

PRIVACY LAWS The two laws most relevant to privacy preference signals are the General Data Protection Regulation (GDPR) and the California Consumer Privacy Act (CCPA). Article 6 of the GDPR [24] establishes a number of legal bases for processing personal data, of which (opt-in) consent is the most common legal basis claimed in a sample of hundreds of AdTech vendors [16, Fig. 2]. Article 4 of the GDPR [24] defines consent as any freely given, specific, informed and unambiguous indication of the data subject's wishes.

Taking a different approach to opt-in consent, the CCPA establishes "the right to direct a business to not sell consumers' personal information" [26, p. 15]. For the purposes of this paper, it is important to note that both laws link the legality of data processing to the privacy preferences of users, creating a need for signals that communicate preferences. However, neither law anticipates the ambiguity resulting from one user sending multiple signals with differing semantics.This is made possible by the decentralised Web ecosystem in which multiple actors and institutions design and propose privacy preference signals, which is described in the following.

PRIVACY PREFERENCE SIGNALS Hils et al. [9] identify five signals that have been adopted at various points in the last 20 years. We ignore P3P because it was deprecated in 2017 and ignore NAI opt-outs as they remain an unpopular and narrow signal [5]. We focus on the remaining three signals. DNT and GPC have a similar technical design in that they extend HTTP headers by a single bit signal, but they differ in semantics. The law does not require recipients to respect DNT, and many ad-tech companies in fact decided to ignore the signal [26, p. 15]. Nonetheless, it can still be turned on in Chrome's and Firefox's settings dialog. In contrast, the Global Privacy Control is designed to trigger the "Do Not Sell" clause (the aforementioned legal right [26, p. 15]) under the CCPA. It also provides a possible interpretation under the GDPR in its specification. However, major browsers have not adopted GPC yet outside of browser extensions.

The third signal, the TCF, is collected via dialogs embedded in the webpage. TCF signals can only be collected by registered intermediaries, of which QuantCast and OneTrust are the most popular [9]. The semantics of this signal are much more complex [16, 22] but revolve around opting-in to various data processing purposes where consent is required. We discuss the nature of the ambiguity resulting from sending DNT/GPC opt-outs and TCF opt-ins in Section 7.5.

Finally, it is worth noting that niche and emerging signals exist that we did not consider. Do Not Sell signals can also be collected via webpages and stored as cookies, which use the standardized *US Privacy String* format [12].

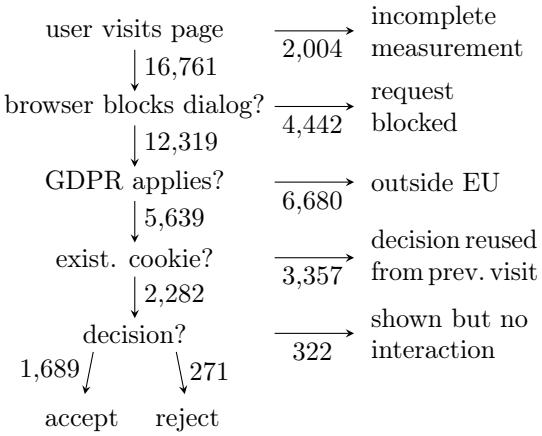


Figure 7.1: Our Consent Dialog Measurement Pipeline

These cookies were successfully reset using the OptMeowt add-on for 17 of 30 websites in a recent study [26]. A technical specification for the Advanced Data Protection Control [10] was proposed that could automatically send privacy preference signals including TCF and Do Not Sell, but does not define any new signals in terms of semantics.

7.3 METHOD

To examine the interplay between privacy preference signals, we embedded Quantcast's cookie consent dialog on the landing page of three websites for a short period of time and also logged visitors' DNT/GPC headers. In constrast to previous research [8], we not only measure a user's decision when they are presented with a TCF consent dialog, but also if they are shown a consent dialog at all. This is important as our findings indicate that a non-negligible number of users employ techniques that block popular consent dialogs entirely.

STUDY PARTICIPANTS We sampled a very technical audience on all three websites. The majority of our measurements were made on mitmproxy.org, the website of an open-source program primarily used by software developers (72% of all impressions) [4]. Additionally, our research group's website and the website of a Capture The Flag contest we hosted contributed 14% of observations each. Note that all numbers in this paper are reported as impressions of the landing page. We do not perform any additional grouping to not overrepresent users who employ additional privacy measures (such as clearing cookies). In total, we observe 16,761 impressions by 8,033 IPv4 addresses from 7,432 /24 subnets.

In terms of browsers used, 53% of visitors used Chrome/Chromium, 23% Firefox, 9% mobile browsers, 8% Safari, and 4% Edge. For comparison, Wikimedia reports 55% Chrome, 13% Firefox, 10% Safari, and 8% Edge on their desktop sites (June 2021).

DATA COLLECTION During our study period we embedded a logging script on all three websites and recorded the following data (see Figure 7.1):

1. the user's browser and whether they sent a GPC/DNT header;
2. state transitions from the browser's page visibility API;
3. whether Quantcast's dialog could be loaded.

If loading the dialog was successful, we additionally recorded:

4. Quantcast's assessment of whether GDPR applies to the current user;
5. existing consent decisions from previous visits;
6. the user's (new) consent decision when a dialog is shown.

To determine whether Quantcast's dialog can be loaded, we manually inject its main script tag and add `load` and `error` event listeners. We cannot detect what prevents this resource from loading, but we suspect that the majority is via ad-blocking browser extensions and DNS-level content blockers.

If Quantcast's JavaScript code could be loaded successfully, we interact with their implementation of the TCF API to determine if GDPR applies to the current user. We only show dialogs to users in the EU, which is Quantcast's default setting.

If we find that GDPR applies, we present the user with a dialog (see Figure 7.3) unless a decision has already been made. We track prior decisions via a cookie set by the dialog. This means the individuals behind the decisions are unique to the extent these cookies are preserved. In total, a consent dialog is only shown for 14% of impressions.

To get a more accurate picture of the user's interaction with the web page, we observe the browser's page visibility API, which emits `visibilitychange` events when the page becomes visible, hidden, or closed. We use the Beacon API (`navigator.sendBeacon`) to record all events as this interface still works when the page is closed. As of 2021, both APIs are available in all major browsers. We discard all measurements for which we did not receive a complete set of events. This also reduces the impact of general network problems, which otherwise may be misattributed as blocking.

RESEARCH ETHICS Our research design requires ethical consideration as each dialog takes user time. Given thousands of websites already impose these dialogs on users [15], we judge that the time cost is outweighed by the value of information derived from our study in shaping the design and adoption of these dialogs. Previous research has shown that Quantcast's dialogs are completed in 3.2s on average [8]. While our institution does not require IRB review for minimal risk studies, we ensured that we did not deceive or harm website visitors and their privacy. All displayed consent notices functioned as described and respected the visitor's choice.
Explaining the research purpose before/after the experiment would lead to a much longer interruption than the initial dialog.

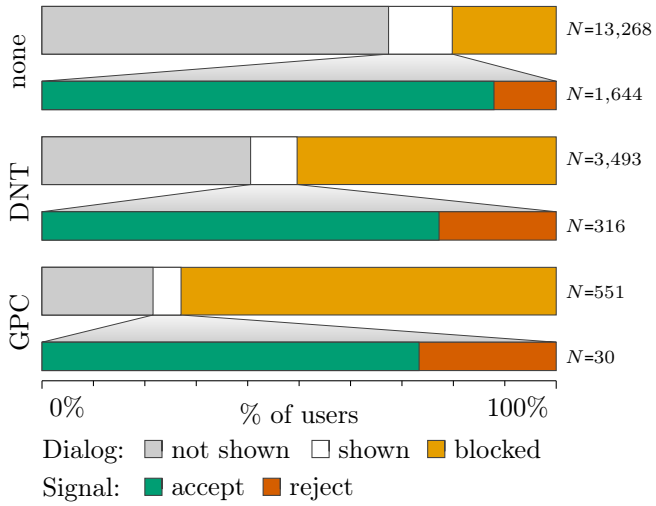


Figure 7.2: The majority of users who have GPC turned on (bottom pair of bars) outright block consent dialogs. Those who do not block are relatively more likely to reject tracking when presented with a dialog offering equal choice. Still, the majority of GPC users click "Accept" and thus send an ambiguous signal. The top pair of bars show the baseline without browser-based privacy signal. The deprecated DNT signal (middle) was more prevalent than GPC, exhibited less extreme blocking, and a similar behavioural response.

7.4 RESULTS

We split our results into two aspects of ambiguity, namely blocked dialogs and multiple signals, and then consider robustness.

BLOCKED DIALOGS The top bars in Figure 7.2 show that collecting a TCF signal via a dialog is non-trivial. The default version of the market-leading dialog does not load for 27% of users, which rises to 50%/73% of users with DNT/GPC enabled. This technical response—blocking the privacy preference communication channel—was not previously considered in the literature.

MULTIPLE SIGNALS The second source of ambiguity results from users sending multiple signals simultaneously. First, 3.5k/550 of impressions send a DNT/GPC signal respectively. All but three of the GPC impressions also send a DNT signal so we do not further differentiate. Displaying a TCF dialog to these impressions creates the potential for a conflict, namely when users send an accept TCF signal while also sending a DNT/GPC signal. Such a conflict occured for 5% of all impressions, or 77%/73% when looking at DNT/GPC-enabled users only. We discuss the nature of the ambiguity later in the paper.

The co-existence of these signals does not only increase ambiguity as the DNT/GPC signals have explanatory power over privacy preferences expressed via the TCF signal. Users with DNT/GPC enabled were 1.9/2.0 times more likely to click "I do not accept" on the TCF dialog than those without. These results are significant at the $p < 0.01$ level for DNT and $p < 0.05$ for GPC.

Note we cannot reject the null hypothesis that DNT/GPC are drawn from the same distribution ($\chi^2(1, N = 346) = 0.07, p = 0.80$).

ROBUSTNESS We run a number of checks to reduce the risk of spurious findings. It could be that GPC adoption was driven by browsers and browser extensions turning it on by default, as it was done by Brave browser [23]. In our sample, Firefox users are most likely to send a GPC signal (8.9 %), followed by Chromium (2.3 %) and Edge (1.6 %). The share of GPC signals from other user agents is statistically zero. Note that Chromium includes Chrome, the most popular browser on the web, as well as Brave, a niche browser catering pro-privacy and cryptocoin-savy users, which identifies as Chrome in the user-agent string. To our knowledge, at the time of our study Brave was the only browser that sent GPC signals by default without asking the user. While we suspect that a number of the Chromium cases with GPC turned on do originate from Brave, the fact that other browsers emit more GPC signals, in both relative and absolute numbers, reassures us that the results are not purely driven by a single browser's default setting. The finding that users who emit DNT or GPC signals tend to chose more privacy-minded TCF options corroborates the behavioral interpretation.

7.5 DISCUSSION

We discuss how to collect preferences, the nature of the ambiguity, and the validity of our results.

COLLECTING PREFERENCES Research into GDPR cookie consent dialogs consistently shows that dialogs contain dark patterns that erode user autonomy [3, 1, 25, 14, 13, 19, 20, 7]. Our findings could be interpreted as further evidence that industry standard dialogs lead users to express untrue preferences. For example, 77% of DNT-enabled users accept data processing in a TCF dialog despite sending a global "Do Not Track" signal.

On the other hand, one could argue that the browser-controlled signals do not capture the true preferences of users. For example, Brave browser turns the signal on by default and does not provide an off-toggle [23]. The lack of an off-toggle goes even further than Microsoft's decision to turn DNT on by default in 2012, which led an AdTech industry group to withdraw from the DNT initiative [11]. This argument is undermined by new counter-arguments found in the the CCPA, which establishes that "affirmatively choosing products or services with privacy-protective features... is considered a sufficiently clear manifestation of opting out" [26].

This argument may not apply in jurisdictions outside of California. Consider that general purpose technologies like browsers contain multiple features. Users may adopt such technologies for any combination of these features, and so the browser may not have been chosen for the privacy features. For example, in addition to Brave containing "privacy-protective features", it also

contains features beyond privacy protection (e.g., cryptotokens) that change the distribution of advertising revenues. Products and services with a narrow range of privacy relevant features, such as browser add-ons like Privacy Badger, may be more reliable indicators of a user's underlying privacy preferences.

Regardless of how the GPC signal is set, we have shown it explains a significant portion of the variance in expressed privacy preferences, which motivates further research into permanently storing privacy preferences in the browser. In particular, browsers could think about collecting more than 1 bit from users (0 bits in the case of Brave) given this information must apply across a range of jurisdictions. Taking Colorado's new privacy law as an example, the 1 bit GPC signal has to cover not only the "Do Not Sell" clause but also opt-in consent for storing sensitive data. While arguments can be made for how to interpret GPC's single bit under each law, the ambiguity could be used by AdTech firms to interpret the signal in their own interest or even request additional signals to 'clarify' the situtation, which imposes yet more decision burden on users.

Our findings suggest an alternative way forward to protect user privacy, at least for opt-in consent. Unlike opt-out signals which default to tracking, opt-in requirements force data processors to collect a privacy preference signal before processing personal data. If browsers block the interface collecting preferences, such as dialogs embedded in web-pages, then firms have no legal basis under an opt-in requirement. This is particularly relevant given a back-of-the-envelope calculation reveals that users have already wasted at least 2,500 years [9] sending TCF signals. This will likely invoke counter-measures from websites leading to an arms race [18].

AMBIGUOUS SIGNALS Two sources of ambiguity we identified, namely blocking dialogs and multiple signals, are particularly relevant for the GDPR where many data processors [8] rely on an "unambiguous" [Art. 4][24] opt-in consent as a legal basis. Blocked dialogs should be resolved in favour of the user not having provided consent. Multiple signals are less easily resolved and may require focusing on the semantics of each signal. For example, the DNT signal "represents a superset of what is covered by Do Not Sell" [26, p. 15] and so is a less ambiguous objection to some of the TCF's data processing purposes.

While many resolution approaches are conceivable (e.g., most/least privacy-minded, temporal or normative order, user intervention), we defer this question to legal analysis with the Blogosphere providing preliminary arguments [2, 21]. The same sources of ambiguity will likely be resolved differently in non-EU jurisdictions. This creates a technical problem for firms processing data from users across multiple jurisdictions—firms must infer each user's jurisdiction. Could privacy aware users masquerade as residents of the jurisdiction with the strongest privacy rights?

Going beyond ambiguity, users can send two conflicting signals expressed under the same standard when both the webpage and the browser collect

preferences. For example, this could occur with the proposed ADPC signal [10] if a user clicked "I accept" in the TCF consent dialog while sending a "No consent under TCF" ADPC header.

VALIDITY We argue that our study has high ecological validity in that we chose the most popular dialog on the Web [9] and displayed it to users browsing a real website. Unfortunately, we could only present this design to the users of three websites, all of which likely over-sample privacy aware and technically literate users. The majority of our participants visited the website of an open-source program targeting developers, and the remaining from our research group's webpages. Future work could embed the same study in a broader range of websites. In fact, our 'experiment' could be carried out passively by any website collecting TCF signals.

7.6 CONCLUSION

We present first evidence that websites do receive ambiguous privacy preference signals, namely opt-in TCF signals sent alongside a GPC opt-out signal. Moreover, the share of ambiguous privacy signals due to blocked TCF dialogs is significant. Both phenomena have been overlooked in the empirical literature. Finally, our study suggests that user adoption of the GPC helps to explain privacy preferences, and is associated with a greater propensity to reject consent.

REFERENCES

[1] Idris Adjerid, Alessandro Acquisti, Laura Brandimarte, and George Loewenstein. 2013. Sleights of Privacy: Framing, Disclosures, and the Limits of Transparency. In *Proceedings of the Ninth Symposium on Usable Privacy and Security (SOUPS '13)*. ACM, Article 9, 11 pages. https://doi.org/10.1145/2501604.2501613

[2] Robin Berjon. 2021. *GPC under the GDPR*. https://berjon.com/gpc-under-the-gdpr/

[3] Rainer Böhme and Stefan Köpsell. 2010. Trained to Accept? A Field Experiment on Consent Dialogs. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '10)*. ACM, 2403–2406. https://doi.org/10.1145/1753326.1753689

[4] Aldo Cortesi, Maximilian Hils, Thomas Kriechbaumer, and contributors. 2010–. mitmproxy: A free and open source interactive HTTPS proxy. https://mitmproxy.org/

[5] Pam Dixon. 2007. *The Network Advertising Initiative: Failing at Consumer Protection and at Self-Regulation*. World Privacy Forum. http://www.worldprivacyforum.org/wp-content/uploads/2007/11/WPF_NAI_report_Nov2_2007fs.pdf.

[6] Global Privacy Control. 2021. GPC Privacy Browser Signal Now Used by Millions and Honored By Major Publishers. https://globalprivacycontrol.org/press-release/20210128.

[7] Hana Habib, Sarah Pearman, Jiamin Wang, Yixin Zou, Alessandro Acquisti, Lorrie Faith Cranor, Norman Sadeh, and Florian Schaub. 2020. "It's a Scavenger Hunt": Usability of Websites' Opt-Out and Data Deletion Choices. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '20)*. ACM, 12 pages. https://doi.org/10.1145/3313831.3376511

[8] Maximilian Hils, Daniel W Woods, and Rainer Böhme. 2020. Measuring the Emergence of Consent Management on the Web. In *Proceedings of the Internet Measurement Conference 2020 (IMC '20)*. ACM.

[9] Maximilian Hils, Daniel W. Woods, and Rainer Böhme. 2021. Privacy Preference Signals: Past, Present and Future. *Proceedings on Privacy Enhancing Technologies* 4 (2021). https://petsymposium.org/2021/files/papers/issue4/popets-2021-0069.pdf

[10] Soheil Human, Max Schrems, Alan Toner, Gerben, and Ben Wagner. 2021. *Advanced Data Protection Control (ADPC)*. https://www.dataprotectioncontrol.org/adpc-spec/

[11] Interactive Advertising Bureau. 2012. "Do Not Track" set to "On" by Default in Internet Explorer 10—IAB Response. https://www.iab.com/news/do-not-track-set-to-on-by-default-in-internet-explorer-10iab-response/.

[12] Interactive Advertising Bureau. 2020. *IAB CCPA Compliance Framework For Publishers & Technology Companies.* https://iabtechlab.com/standards/ccpa/

[13] Dominique Machuletz and Rainer Böhme. 2020. Multiple Purposes, Multiple Problems: A User Study of Consent Dialogs after GDPR. *Proceedings on Privacy Enhancing Technologies* 2 (2020), 481–498. https://doi.org/10.2478/popets-2020-0037

[14] Arunesh Mathur, Gunes Acar, Michael J Friedman, Elena Lucherini, Jonathan Mayer, Marshini Chetty, and Arvind Narayanan. 2019. Dark patterns at scale: Findings from a crawl of 11K shopping websites. *Proceedings of the ACM on Human-Computer Interaction* 3, CSCW (2019), 1–32.

[15] Célestin Matte, Nataliia Bielova, and Cristiana Santos. 2020. Do Cookie Banners Respect my Choice? Measuring Legal Compliance of Banners from IAB Europe's Transparency and Consent Framework. In *IEEE Symposium on Security and Privacy*. IEEE, 791–809. https://doi.org/10.1109/SP40000.2020.00076

[16] Célestin Matte, Cristiana Santos, and Nataliia Bielova. 2020. Purposes in IAB Europe's TCF: which legal basis and how are they used by advertisers?. In *Annual Privacy Forum*.

[17] Jonathan Mayer and Arvind Narayanan. 2011. Do not track-universal web tracking opt out. *Center for Internet and Society* (2011).

[18] Rishab Nithyanand, Sheharbano Khattak, Mobin Javed, Narseo Vallina-Rodriguez, Marjan Falahrastegar, Julia E. Powles, Emiliano De Cristofaro, Hamed Haddadi, and Steven J. Murdoch. 2016. Adblocking and Counter Blocking: A Slice of the Arms Race. In *6th USENIX Workshop on Free and Open Communications on the Internet (FOCI 16)*. USENIX Association, Austin, TX. https://www.usenix.org/conference/foci16/workshop-program/presentation/nithyanand

[19] Midas Nouwens, Ilaria Liccardi, Michael Veale, David Karger, and Lalana Kagal. 2020. Dark Patterns after the GDPR: Scraping Consent Pop-Ups and Demonstrating Their Influence. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems (CHI '20)*. ACM, 13 pages. https://doi.org/10.1145/3313831.3376321

[20] Sean O'Connor, Ryan Nurwono, and Eleanor Birrell. 2020. (Un)clear and (In)conspicuous: The right to opt-out of sale under CCPA. arXiv:2009.07884 [cs.CR]

[21] Harshvardhan J. Pandit. 2021. *GPC + GDPR: will it work?* https://harshp.com/research/blog/gpc-gdpr-can-it-work

[22] Cristiana Santos, Nataliia Bielova, and Célestin Matte. 2020. Are cookie banners indeed compliant with the law? Deciphering EU legal requirements on consent and technical means to verify compliance of cookie banners. *Technology and Regulation* (2020), 91–135.

[23] Peter Snyder and Anton Lazarev. 2020. *Global Privacy Control, a new Privacy Standard Proposal, now Available in Brave's Desktop and Android Testing Versions.* https://brave.com/global-privacy-control/

[24] European Union. 2016-05-04. Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation). *Official Journal L110* 59 (2016-05-04), 1–88.

[25] Christine Utz, Martin Degeling, Sascha Fahl, Florian Schaub, and Thorsten Holz. 2019. (Un)informed Consent: Studying GDPR Consent Notices in the Field. In *Proceedings of the 2019 ACM SIGSAC Conference on Computer and Communications Security (CCS '19)*. ACM, 973–990. https://doi.org/10.1145/3319535.3354212

[26] Sebastian Zimmeck and Kuba Alicki. 2020. Standardizing and Implementing Do Not Sell. In *Proceedings of the 19th Workshop on Privacy in the Electronic Society* (Virtual Event, USA) *(WPES'20)*. Association for Computing Machinery, New York, NY, USA, 15–20. https://doi.org/10.1145/3411497.3420224

APPENDIX

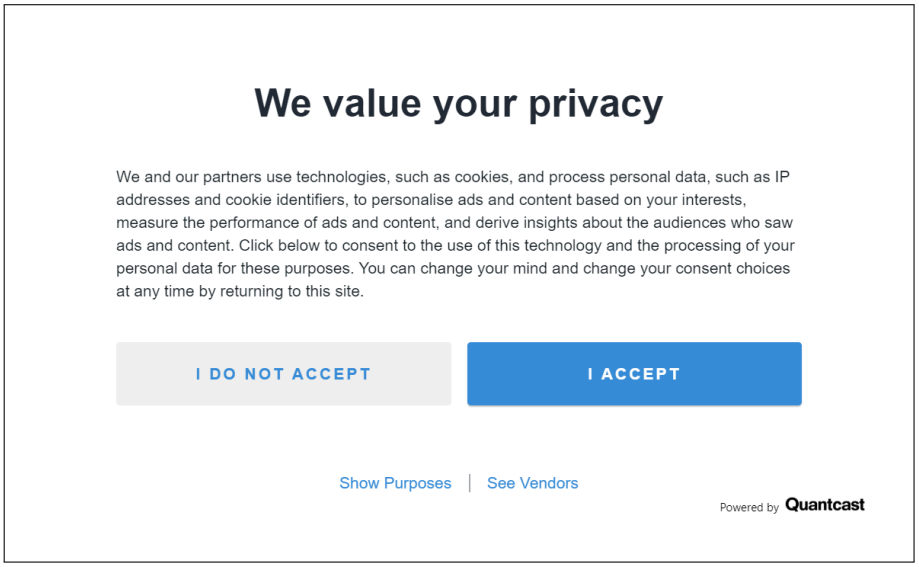


Figure 7.3: Quantcast's TCF cookie dialog used in this study. Note that we used Quantcast's default configuration, which renders the "I accept" button more prominently and thus does not provide an equal choice.