Preserving Needles in the Haystack: A search engine and multi-jurisdictional forensic documentation system for privacy violations on the web

Timothy Libert (corresponding), Anokhy Desai, Dev Patel
public@timlibert.me,anokhyd@andrew.cmu.edu,dnpatel@andrew.cmu.edu
Carnegie Mellon University

ABSTRACT

Numerous companies track the browsing habits of billions of web users, yet their own practices remain opaque. This information asymmetry favors companies to the disadvantage of policy enforcement. To resolve this situation we present a search engine for privacy violations on the web which leverages a database of 1.4 billion HTTP requests, 300 million cookies, and one million policies, derived from 11.5 million stateful page loads across 2.3 million sites.

Our system allows regulators and litigants to find and document highly-specific privacy violations in particular legal jurisdictions so they may take legal action. Violations can be found based on policy and page content, third-party code ownership, and coding implementations. Court-admissible evidence of violations can be collected from measurement nodes located at residential IP addresses in specific countries. In addition to finding violations of extant law, our large measurement corpus can model the impact of potential policies, thereby better informing legislative decisions.

We use our system to reverse-engineer a GDPR violation related to targeting advertisements based on mental health from a residence where GDPR applies. We locate 43 websites directed to children which claim compliance with U.S. children's privacy regulation, yet run afoul of the law by not implementing Facbook's child privacy features. We find implementing GDPR-style protections on political advertising in the U.S. could have significant impact on nine companies performing political ad targeting using sensitive categories, five of whom already explicitly disallow such targeting in the E.U.

CCS CONCEPTS

• **Security and privacy → Privacy protections**; • **Applied computing → Law**; **Evidence collection, storage and analysis**.

KEYWORDS

web, privacy, forensics, law, tracking, children, mental health, policies

1 INTRODUCTION

Decades of research have demonstrated most people find web tracking objectionable and yet it remains widespread. The primary countermeasures to web privacy violations proposed by the research community have been technical, in the form of browser add-ons, and policy-based, in the form of improved summaries, notifications, and consent dialogues. New laws such as GDPR and CCPA have promised enhanced oversight and fines. Despite much potential, these strategies often operate in a poorly-coordinated fashion and measurement studies regularly demonstrate that tracking has not been significantly curtailed.

Technical countermeasures by themselves have failed to have lasting impact as they are almost immediately met with new forms of tracking. While ever-evolving tracking technologies annually beget a new crop of conference papers detailing methods found "in the wild", 20 years of cat-and-mouse games have not arrested the growth of tracking. In the absence of clear legal requirements, improved policy and consent systems require good faith implementations from technology companies. Such faith is unmerited, as companies often benefit financially when users are not able to effectively limit data collection or are misled into accepting practices they may not understand. It can come as little surprise that industry notifications such as the "AdChoices" icon have been found to be insufficient at best [35].

Initial research suggests privacy laws such as the GDPR have impacted non-consensual tracking, yet European Data Protection Authorities (DPA's) have produced a paucity of sizable fines. Some regulators, such as the UK's Information Commissioners Office (ICO), have outright retreated from their enforcement duties, not wanting to put "undue pressure" on web tracking companies.[1] While the adtech industry braced themselves for a tsunami of GDPR fines in 2018, it is the US Federal Trade Commission, which has limited legal authority to regulate privacy, who have produced the most sizable privacy fine to date ($5B against Facebook[2]). Three possible explanations for weak GDPR enforcement are the fact that many companies have resources to fight legal battles around the world, some regulators have lapsed into a state of inaction after initially taking a "wait and see" approach, and fundamental difficulties coping with the complexity of the web tracking ecosystem.

Despite weak enforcement to date, the most effective means to combat online tracking is to apply regulatory and legal powers at a far more aggressive level. The reason for this is fairly intuitive: tracking is done to make money. If the cost of fines, lawsuits, and legal proceedings exceeds the money being made from tracking, it will cease in many forms. Contrary to claims from adtech companies, it is quite unlikely putting a dent in tracking will cause undue harm to publishers of online content, meaning clamping down on targeted advertising will have few negative externalities. For example, research shows tracking-based advertisements give publishers an "average increase of $0.00008 per advertisement" [24], a New York Times executive recently stated that ceasing tracking-based advertising in Europe has not negatively impacted revenue, and in

[1] https://ico.org.uk/about-the-ico/news-and-events/news-and-blogs/2020/05/ico-statement-on-adtech-work/
[2] https://www.ftc.gov/news-events/blogs/business-blog/2019/07/ftcs-5-billion-facebook-settlement-record-breaking-history

fact, the "digital advertising business continues to grow nicely"[3][4], and when a Dutch broadcaster's website "switched from tracking-based targeting to contextual targeting, revenue increased 61%".[5] Thus, while adtech companies present themselves as saviors of the news industry, they serve a middle-man role, extracting data and value from publishers while simultaneously violating user privacy [21]. Users and publishers alike are better off without invasive web tracking.

But how is more aggressive enforcement to be accomplished without giving regulators massive budgets? Our contribution is to combine concepts from technical, policy, and regulatory approaches to put regulators on equal footing with companies. We use state-of-the-art web measurement to create a massive searchable database of tracking mechanisms, with 1.4 billion HTTP requests, 300 million cookies, and one million policies, derived from 11.5 million stateful page loads across 2.3 million sites. We harvest a broad array of policy types (Privacy Policies, Terms of Service, Cookie Notices, Ad Choices, GDPR Statements, and CCPA Statements) from sites and third-party companies on a heretofore unseen scale. Finally, we draw lessons from law and computer forensics to create a forensic documentation system that generates evidence in specific legal jurisdictions on demand, allowing regulators and litigants to easily document privacy violations in a court-admissible fashion.

This paper has five main aims. First, we detail our methodology which utilizes a distributed architecture to achieve a high data ingest rate for forensic-quality scans as well as performing geographically-targeted remote measurements. Second, we share optimal strategies for surfacing trackers using a stateful intra-site crawl strategy that we run against a population of 2.3 million websites, avoiding many issues associated with common top-site lists. Third, we present macro-level trends in tracking and policies which update and expand upon prior studies. Fourth, we demonstrate how our tool can find incredibly specific, and potentially legally actionable, privacy violations in the domains of mental health and children's privacy. Fifth, we show how changes to federal law could impact online political advertising in the United States. Our discussion focuses on how web tracking companies have grown accustomed to profiting from *user surveillance* and the implications of applying *regulatory surveillance* to these companies. It is our hope that the many deleterious impacts of surveillance felt by users will soon be felt by tracking companies.

2 BACKGROUND: PRIVACY VIOLATIONS AND REDRESS MEASURES ON THE WEB

Our work extends and synthesizes many prior areas of web privacy research. We are motivated by long-standing research showing that most people find web tracking objectionable yet widely prevalent. While we find benefits coming both from technical and policy approaches to web privacy, a lack of synergy means neither approach is reaching its full potential. Finally, while some initial positive effects of new privacy regulations exist, we find fines are falling far short of what is possible, and tracking has not ceased.

Privacy has been found to be a near-universal concern, both in regards to government and companies. A 2019 Pew Research Center study showed 81% of U.S. adults feel that they lack control over their private information and 79% have concerns about their data being collected by companies [2]. In 2018, the U.S. National Telecommunications and Information Administration found that most Americans have privacy and security concerns [14]. Further studies have shown privacy concerns and fear of government surveillance has had a "chilling effect" on people seeking out news, leading them less likely to comment on controversial topics [25, 33].

Despite public concerns over the practice, privacy violations are endemic to the web. This is largely due to the inclusion of third-party code by websites, which is used for optimizing website design, social media sharing, audience measurement, marketing, and more [21]. When third-party code is included in a site, it often exposes users' browsing habits to companies hosting the code, and may be further augmented with persistent identifiers such as cookies [8, 19], DOM storage [12, 30], and fingerprinting [3, 12, 13]. Tracking has likewise been found on smartphone applications [37, 39] and in physical retail locations [34]. Personal data collected by one company is often shared with others. Most famously, Facebook exposed the data of 87 million users to Cambridge Analytica, a firm attempting to use the data for political manipulation [15].

Technical responses to web tracking have primarily been aimed at stopping tracking via browser improvements and extensions. Privacy Badger, Disconnect, and Ghostery are some of the more popular privacy extensions [7][6], with the latter two relying on a block-list approach that requires active maintenance. Studies have shown that Ghostery and Disconnect are efficient tools in blocking third-party trackers, but browser extensions still have blind spots that allow tracking [28]. Some browser vendors like Apple are implementing features without the need for extensions (e.g. Intelligent Tracking Prevention[7]), but adtech companies are not only finding work-arounds, but openly sharing their techniques [23].

An additional issue limiting the efficacy of browser add-ons is usability. Balebako et al. conducted a study to investigate 9 widely used privacy tools and found that most were difficult to configure and led participants to think that blocking was functional when it wasn't [18]. Ghostery, despite being generally effective, uses jargon that is confusing to users [18]. Finally, blocking trackers often breaks page layout and users may disable privacy tools in response [26].

In addition to blocking tracking via technological solutions, others have sought to help users make privacy choices based on existing policies. Numerous studies have shown that the current privacy policy-based system is flawed due to difficulties in understanding lengthy and complex documents [20, 27]. To cope with the complexity of policies, some have proposed a "nutrition label" approach by which policies are summarized using standard layouts [11, 17, 31], while others have have suggested policy summarization and privacy icons [6, 10, 38]. While these approaches attempt to make legal documents comprehensible to users, they are limited by a lack of enforceable requirements and high incentives for companies

[3] https://digiday.com/media/gumgumtest-new-york-times-gdpr-cut-off-ad-exchanges-europe-ad-revenue/
[4] The Times is going so far as to support a "Do Not Track" successor: https://globalprivacycontrol.github.io/gpc-spec/.
[5] https://brave.com/npo/
[6] https://www.ghostery.com/
[7] https://webkit.org/blog/9521/intelligent-tracking-prevention-2-3/

to stymie user attempts at making privacy choices. Others take a different approach, analyzing policies to find areas to hold companies and sites accountable for violations [20, 39, 40], which we find under-explored, and more promising.

In an attempt to bring the power of the law to bear on the problem, the 2018 European Union General Data Protection Regulation (GDPR) has sought to put clear limits on processing personal information, backed by steep fines. Initial findings suggest the threat of fines has reduced the amount of non-consensual tracking and encouraged companies to put Privacy Policies on their websites [8, 9, 22]. However, other research has shown the improvements to be more modest [32, 36]. Johnson et al have shown tracking rising after an initial drop when the law came into effect, with the unintended consequence of increased concentration in the tracking market, thereby benefiting large companies [16]. Although GDPR allows fines up to 4% of global revenue, which would amount to nearly €2.8 billion for a company like Facebook, [8] European regulators have thus far levied a paltry €554 million in fines *total*.[9] eMarketer estimated the size of the global digital ad market, much of which is powered by tracking, at $335 billion (€285 billion) in 2019. Put in perspective, all GDPR fines to date represent 1/570th of the total online ad market, hardly an existential threat. Given the highly lucrative nature of digital advertising relative to fines levied, GDPR has so far been more bark than bite.

While we see merit in all the approaches detailed above, we believe potential synergies are being missed. First, measurement studies using platforms such as OpenWPM and webXray have been oriented towards one-off single-point censuses intended for research audiences rather than providing robust search tools and distributed in-jurisdiction measurements [12, 19]. Second, technical blocking approaches perform well in many, but not all, cases, and many insights and techniques could be used to document and prove violations in a legal setting as well. While policy summaries can theoretically help some users, they could be more useful if policy data were instead made more digestible for regulators and litigants. Finally, if the above approaches are brought together in service of enforcing policy, better regulatory outcomes could be achieved. This requires a fundamental change in mindset: moving away from viewing tracking as a problem to be "solved," and towards treating the web like a crime scene to be forensically documented.

3 LEGAL EVIDENCE AND THE WEB

To take legal action against companies violating web user privacy, evidence of privacy violations must be submitted in court. We herein consider how web measurement data relates to the four traditional types of evidence: real, demonstrative, documentary, and testimonial.[10] Real evidence is a tangible item that must be relevant and authentic, like a knife or blood sample. Demonstrative evidence, like a chart or diagram, illustrates witness testimony in a digestible manner. Documentary evidence, like letters and contracts, is an authentic and trustworthy document that can prove or disprove an allegation. Testimonial evidence is given by a witness under oath about the facts of a case.

To be admitted into court, evidence must be relevant, material, and component, meaning it must help prove or disprove a fact. For example, a site's source code is admissible if it proves a fact necessary to make an argument, such as a tracking cookie being set. Evidence authenticity must be proven through methods like expert witness testimony, chain of custody documentation, or other proof of distinctive characteristics of that piece of evidence. "Other proof" includes the use of file hashes such as MD5 or SHA, which typically meet the legal prerequisites for admissible digital evidence.

Technical data and digital artifacts can fit into more than one category of evidence. In ongoing Facebook litigation, plaintiffs have brought in technical evidence like JavaScript code, AdClick data, and the company's Privacy Policies, with expert witnesses deposed or asked to testify to explain the evidence.[11] The Wayback Machine, a digital archive of the web, has been frequently used in intellectual property cases to determine issues of infringement or patent, trademark, and copyright invalidation. Similarly, there is precedent in intellectual property and patent infringement cases for the admissibility of software components like a program's code or hashing function.[12] Code and analytics data can be collected on a hard drive and examined by forensics professionals, who then testify about their findings and speak to the data's authenticity, making the data both demonstrative and testimonial. Printed screenshots of a database or Privacy Policy would be considered documentary evidence.

Evidence that is stored or transmitted in digital form can be used at trial after it is determined that it is relevant, authentic, and not hearsay [5]. Over the last few decades, digital evidence has increasingly been admitted in the form of emails, digital photographs, chat and browser histories, spreadsheets, databases, computer memory logs and backups, and video/audio files. While digital evidence is often attacked for its authenticity because it can easily be modified (eg by changing the creation date of a Word document), there has been a wider admissibility of purely computer-generated data, such as server logs, due to the fact that there is less direct human interaction in their generation and less doubt regarding authenticity. Software itself has been inherently trusted by most courts because it is viewed as computer-generated evidence [4]. Network traces and internal browser state, which is what we focus on, may be considered computer-generated evidence as well.

4 RESEARCH QUESTIONS

The underlying motivation for this work is to determine if combining approaches from web measurement, technical analysis, and policy analysis can be used to locate and forensically document privacy violations in specific legal jurisdictions at scale. There are two primary benefits from this approach. First, regulators and litigants may be empowered to pursue legal action. Second, we may model the impact of proposed policy changes on extant tracking behaviors, so new regulations are informed by evidence. Several research questions help us evaluate if our approach is sound.

[8] https://www.statista.com/statistics/268604/annual-revenue-of-facebook/
[9] https://www.coreview.com/blog/alpin-gdpr-fines-list
[10] https://corporate.findlaw.com/litigation-disputes/summary-of-the-rules-of-evidence.html

[11] In re Facebook Privacy Litigation, 192 F.Supp. 2d 1053 (N.D. Cal. 2016)
[12] Finjan, Inc. v. Blue Coat Systems Inc., 879 F.3d 1299 (Fed. Cir. 2018)

4.1 Results of Large-Scale Stateful Crawl

We believe our dataset to be one of the largest stateful crawls presented in academic literature thus far. Likewise, we collect an array of policies from websites and third-parties which have not been studied at this scale. Several questions flow from this:

RQ1: What are current macro-level tracking trends?

RQ2: What are macro-level features of different types of policies? Is there a correlation between the presence of a policy and amount of tracking?

4.2 Enforcing Extant Regulation

As noted above, digital evidence provided to a court must prove a given fact, such as how the source code of a page results in a privacy violation. We use our tool to pursue two case studies by forensically documenting privacy violations at the code level, answering the following questions:

RQ3: Can we reverse-engineer the process by which cookies are set on a mental health site in a GDPR-covered jurisdiction by a company that targets ads based on mental health?

RQ4: Can we use website policies as a guide to isolate and identify failures in client-side tracking implementations? Specifically, can we find websites for children which claim to comply with the Children's Online Privacy Protection Act (COPPA), but do not properly implement Facebook's child-safe privacy controls?

4.3 Modeling Impact of Proposed Regulations

In addition to enforcing laws that exist, it is helpful to know the potential impact of a new or proposed law. There are three main values for this approach: first, using independent evidence can help make the policy process more impartial and less susceptible to special interests; second, laws may be precisely tailored to specific objectionable practices; and third, knowing practices in advance may give regulators a head start on compliance actions.

RQ5: Can applying GDPR's sensitivity guidelines identify areas where new regulation on political advertising in the United States may have an observable impact?

5 METHODS

Our goal is to create a search engine to find and forensically document privacy violations, and our methods address several main issues: the type of browsing data to make searchable, the type of policies to collect, how to connect tracking to the companies responsible, how to specify sites real people are likely to visit, how to massively scale measurement, and how to perform in-jurisdiction measurements.

5.1 Web Measurement Data Collected

We want to model privacy violations experienced by real users, so we use a production version of the Chrome desktop browser, the same browser that roughly 70% of people use worldwide, and develop a crawl technique optimized to surface the most trackers.[13] This factor alone offers a significant improvement over OpenWPM, which uses Firefox, a browser with an uncertain future used by fewer than one in ten people.[14]

We have a customized browser instrumentation by which we use Selenium primarily for launching browser instances, and conduct nearly all of our automation using Chrome's Devtools Protocol (CDP) via a websocket connection.[15] CDP has a rich API for browser interaction, and does not require injecting scripts into the page to accomplish most tasks. This leaves the browsing session free of external modifications, which could bring into question the accuracy of measurements. We use CDP to extract the content of the page, including scripts, source code, and binaries, as well as page behaviors like network activity (requests, responses, websockets, event source messages), session storage (cookies, DOM storage), security certificates, hyperlinks, and more. *After* we have made measurements of the page, we optionally inject the Mozilla-maintained Readability.js library to extract text.[16] This process is self-contained in a client and returns a JSON object to the calling system, a fact we will return to when we describe our distributed architecture.

All of the above data types are selectable depending on needs. For example, a scan can be done just to retrieve network data, or to exclude the content of response bodies. The reason for this is we conduct two types of scans: "haystack" and "forensic". Haystack scans are intended to be done at scale on arbitrary websites, and thus do not require the collection or storage of binaries. In contrast, forensic scans are designed to capture as much as possible, so it may be recorded as evidence. This distinction allows us to achieve massive scale with haystack scans, which in turn allows us to find targets for forensic scans. We also conduct extensive pre-processing once scan data is returned from the client, which has two benefits: first, we reduce overhead and dependencies for the browser instrument; second, we speed up search tasks later on. Additionally, code and binary data are hashed to speed up queries, reduce disk space, and provide a forensic record the contents have not been tampered with. We record all events with millisecond resolution to provide evidentiary documentation and facilitate reverse-engineering.

5.2 Policies Collected

In regards to policies, we want to capture the entire range of legal agreements users are subject to, not just Privacy Policies, as has been done in the past. We collect Privacy Policies, Terms of Service, Cookie Policies, Ad Choices Policies, CCPA Statements, and GDPR Statements. To locate policies, we first use an inclusive approach to flag URLs that *could* be a policy based on heuristics such as including terms like "privacy" or "terms". Next, we visit the page and extract the title and text. We then use another set of heuristics such as the phrases "Privacy Policy" and "Terms of Service" to determine if a page is a policy, starting by examining the title, and then moving to header tags, bold text, and then normal text. We randomize the order we check for policy types, thus a page titled "Privacy Policy and Terms of Service" has an equal chance of being assigned either "Privacy Policy" or "Terms of Service". We choose a single-type assignment strategy rather than multi-layer classification to

[13] https://gs.statcounter.com/browser-market-share/desktop/worldwide

[14] To be clear, we don't celebrate this fact, but are compelled to point it out from a methodological perspective.

[15] https://chromedevtools.github.io/devtools-protocol/

[16] https://github.com/mozilla/readability

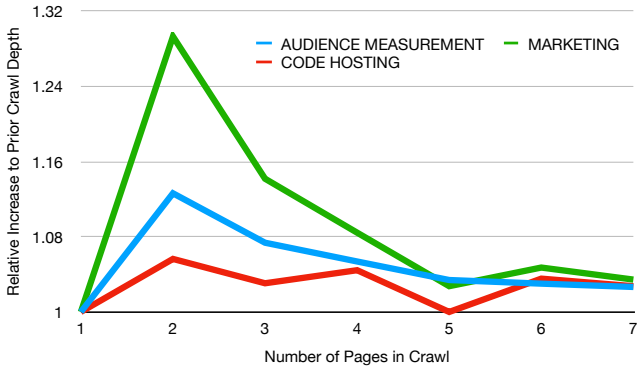

Figure 1: Visiting more pages loads more third-parties relative to a lower crawl depth, with changes in marketing domains (green) as the biggest factor overall, while code hosting (red) and audience measurement (blue) show less variation.

simplify our analysis. We currently support classifying policies in five languages (English, German, French, Spanish, and Chinese).

5.3 Optimal Crawl Strategy

Our browsing instrument has three different modes: single page scan, crawling a pre-defined page list, and crawling random pages within the same site. Each has benefits: single page scans are fastest, pre-defined lists allow replay of browsing sessions, and random crawls provide many pages from the same site with no foreknowledge of site structure. For haystack scans, we find random crawls are preferable, as visiting many pages is likely to result in more tracking detected. However more pages come with a time cost, thus negatively impacting speed, and ultimately, scale. We conducted an experiment by which we randomly crawled between 1 and 7 pages for roughly 4,000 sites to find the optimal strategy.

The results of our experiment revealed significant changes in the amount of third-party domains loaded per crawl. A homepage-only crawl found an average of 30 domains per site, whereas a 2-page crawl found 40, and a 7-page crawl found 58. To determine the point at which returns began to diminish, we calculated the increase in domains detected in each crawl *relative* to the one before (see Figure 1). We found that while there is a significant jump when the second page is visited, increased benefits stabilize after five pages. Likewise, marketing code is the biggest factor, and we hypothesize this may be due to sites interpreting multiple page views as the client not being a bot or as a form of implicit consent to be tracked (e.g. "by using this site you consent to..."). Given trade-offs between speed, stability, and depth, we find crawling 5 pages per site to be the optimal strategy, **giving us a roughly 1.75x improvement in third-party detection over homepage-only studies** [12, 19]. We also have several optimizations to allow sites more time to load resources if needed, or to return results faster if network activity ceases, which increases speed without sacrificing accuracy, thereby enhancing scale.

5.4 Tracking Attribution to Companies

In order to take legal action, it is necessary to know who is responsible for a given privacy violation: evidence of a crime with no culprit has little value. A key component of our method is our domain owner list which links domain names (e.g. "yimg.com") to their owners (e.g. "Yahoo!") as well as any parent companies (e.g. "Verizon"). Our list is loosely based on data originally released as the webXray Domain Owner List [21], but has been expanded over a number of years to cover 2,986 domains, 813 companies, and 4,227 policies in 70 languages.

When we find third-party domains not in our list, we manually seek out the owner, and in turn, any parent companies. Our first step when encountering a new domain is to reference *whois* data. **However, we find that 36% of domains in our dataset have anonymous *whois* registration, indicating many companies actively seek to avoid attribution of their tracking activities.** In cases where *whois* is insufficient, we use a variety of methods, including web search, to find owners. Crunchbase is our primary source of parent company data.

In addition to ownership, we locate the homepages for each company, Privacy Polices, Terms of Service, GDPR Statements, Cookie Policies, Ad Choices Policies, CCPA Statements, and Opt-Out URLs. We collect policies found in any language and store them with the appropriate language code. We classify domain owners by the types of platforms they collect data from (e.g. "web," "mobile," "TV," and "IoT")[17], type of service the company provides (e.g. "hosting," "marketing," "fonts"), the country the company is based in (e.g. "US," "CN"), and industry trade group memberships (e.g. "NAI," "DAA").

The main uses of our list are for tracking attribution (e.g. connecting a cookie or data transfer to the entity which owns a domain), searching for companies engaged in certain activities (e.g. hosting fonts), or making certain policy claims (e.g. not allowing political advertising in the EU). For example, we can easily find all Privacy Policies of companies in the Network Advertising Initiative (NAI) containing the words "political" or "health", a capability we leverage for our case studies.

5.5 Site Selection Methodology

Our goal is to measure browser behavior on websites real users are likely to visit and interact with. Prior web measurement studies have relied on "top site" lists of dubious quality for site populations. [29] Common problems with such populations are opaque criteria for list inclusion, susceptibility to manipulation, sites which are merely landing pages or "domain for sale" placeholders rather than sites real people would visit, and URLs which redirect to different domains entirely. The Tranco list attempts to resolve manipulation problems by combining data from the Alexa, Umbrella, and Majestic lists. [29] However, the primary Tranco list is limited to one million sites and provides no guarantees the sites would be visited by a real user or that domains do not redirect. The list thus does not fulfill our needs.

To build our site population, we utilize the Chrome User Experience Report (CrUX) which "provides user experience metrics for how real-world Chrome users experience popular destinations on the web" using data "aggregated from users who have opted-in to

[17] Mobile, TV, and IoT data is derived from datasets not covered herein.

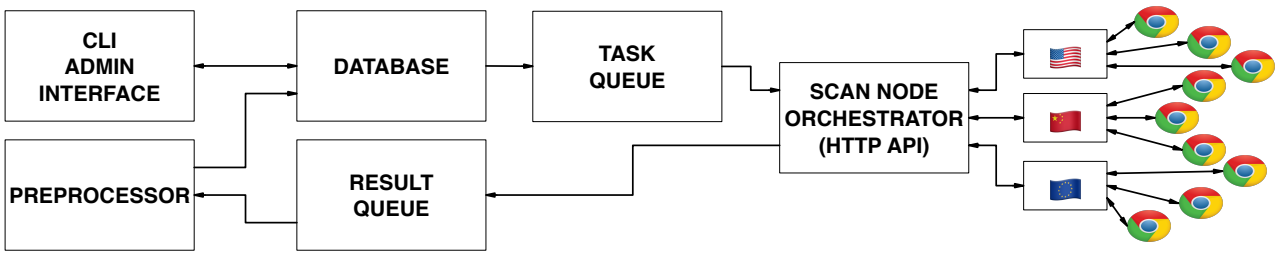

Figure 2: High-level view of our distributed architecture: The system runs off a PostgreSQL database from which a node orchestrator retrieves tasks. The orchestrator assigns tasks to nodes based on geographic location. When the orchestrator receives a finished task the result is stored in a queue and then preprocessed.

syncing their browsing history"[18]. Given that we utilize Chrome for measurements and it is the most popular browser in the world, CrUX is best suited to our needs. To ensure that the sites we examine are consistently visited by real people, we identify sites which occur in the CrUX global dataset consistently for six months. Upon loading a given URL, we discard websites that redirect to another domain or have fewer than five links to the same site [19] so we may avoid "domain for sale" pages and the like. Based on this method, we attempted scans of over three million sites and completed acceptable scans for 2.3 million.

5.6 Distributed Infrastructure

As discussed above, numerous design considerations have been made to ensure that each component of our system is as fast as possible. However, to truly achieve scale, we need to massively parallelize our system with a distributed architecture, which is the key component of our approach. Our current system has throughput of 13k - 30k pages per hour depending on configuration, meaning we can perform upwards of 5 million scans per week. We believe our system can scale up further with relative ease, and view the current data set, however large, as a proof-of-concept.

The best way to describe our architecture is to walk though how a scan is performed (see Figure 2). First, a command-line administration program is used to specify a name for a new database (e.g. "million_site_scan"), the type of task to be performed (e.g. "get_scan"), and a text file with URLs (e.g. "tranco_1m.txt"). The administration program connects to a PostgreSQL database, instantiates a new, highly-structured, relational database and puts tasks into a queue. An orchestration server operates an HTTP-based API which coordinates nodes with specific databases. When a node contacts the orchestration server it matches the node's ID to the assigned database, and sends back the appropriate task. When the node completes the task, it is uploaded back to the orchestration server, which stores the result to a master result queue. Storage nodes remove items from the result queue, performs extensive processing, and store results to the database (more detail below). At this point, results may be queried and reports run.

[18]https://developers.google.com/web/tools/chrome-user-experience-report/

[19]Redirects and link evaluation is based on public suffix+1: "example.co.uk" -> "www.example.co.uk" is not counted as a redirect whereas "example.co.uk" -> "example.net" is.

The scan nodes deserve additional discussion. For our haystack scans, we have roughly 20 machines on a university IP which range in power from quad-core i7 to dual Xeon eight core configurations with at least 1GB of memory per processor core. We use rack-mounted servers, space-saving Intel NUCs, as well as any machines we find around our office not in use. As our only requirement is to run Chrome, even older, mid-spec, desktop computers abandoned from other research projects can be added to the cluster. Each node has a client program that runs one sub-process per processor core. The sub-process polls the API for tasks, launches a Chrome browser, performs a task, compresses the JSON output, and sends it back to the server. The most obvious benefit to this approach is to run several hundred browsers at once at a very low cost. An additional benefit is distributing our tasks across a large range of university IP addresses obfuscates the automated nature of our scans.

Storage nodes take compressed results out of the queue and process the raw data so it may be searched. Some tasks performed include parsing out domains to determine if a request is to a third-party, parsing cookies, parsing and examining request headers, parsing GET and POST data, hashing files and binary data, performing basic analysis of text content, and assigning IDs to crawls based on a hash of the page sequence. Once finished, results are stored to the appropriate database from which the task was taken.

5.7 Multi-Jurisdictional and Residential Measurements

As noted above, we have a cluster of computers at an academic institution for our haystack scans, but for forensic scans, we want to ensure we take complete measures from residences located in a specific area. Our client is written in Python and has only three dependencies: Selenium, Chrome, and ChromeDriver. The client can be easily installed on a laptop or in a virtual machine and a configuration parameter sets the client's ID. At this point, the client can be started as a background task and left to run. For this paper, we have performed forensic scans from residences in the UK and the US, and in our testing, we have successfully deployed on cloud servers in the US, UK, Germany, the Netherlands, and China. Our next goal is to deploy scan nodes all over the world.

5.8 Limitations

Our system is fast and reliable, performs to specification, and is extensively tested. Some of the core code has been in use for nearly

	Privacy	Terms of Service	Ad Choices	Cookie	GDPR Statement	CCPA Statement
N Documents	582,676	338,324	929	80,929	2,874	6,506
% of Sites	26%	15%	<1%	4%	<1%	<1%
Ave. 3Ps for Site w/Policy	27	28	54	26	21	87
Ave. 3Ps for Site w/out Policy	19	19	20	21	21	21
% w/Pacifying Language	32%	13%	21%	10%	33%	26%
% GDPR-Specific Terms	2%	<%1	<%1	<%1	2%	<%1
Ave. Word Count	2168	2668	1266	1362	1707	1669
Ave. Flesch Reading Ease	30	29	31	29	29	32

Table 1: Privacy and Terms of Service are the most prevalent types of policies, whereas others are comparatively rare. Terms of service are the longest policies, followed by Privacy Policies. Sites with policies almost always have more trackers than those without. The reading difficulty is at or near the college graduate level for all policies.

eight years. However, there are features that need to be added and inherent limitations to the approach. We do not currently instrument Javascript sufficiently and are reliant on static code analysis, which we aim to rectify in future versions. While we scroll the page to surface more content and modify our User-Agent for compatibility, we take no additional measures to defeat bot detection, which may skew our findings. An additional point to clarify in regards to forensic data is without a more robust experimental setup we are documenting data *collection* rather than data *use*. However, this may not be a problem as long as litigants and regulators know exactly what to subpoena.

Other problems come from scale: despite pre-processing, search queries can run fairly slowly given the large volume of data. Deeper instabilities in Chrome and Chromedriver can cause the browser to crash and it may take several attempts to get a successful load, or it may fail.[20] The biggest current limitation is the interface is command-line only, and we are exploring developing a web-based version.

6 FINDINGS

We present three case studies which show the power and versatility of our tool when applied at the micro-, meso-, and macro-levels. At the micro-level, we use forensic data collected from a GDPR-covered jurisdiction to reverse-engineer two parties syncing tracking cookies on a mental health website in furtherance of targeting ads based on mental health. At the meso-level, we find 43 websites which claim compliance to U.S. children's privacy law in their polices, yet do not implement Facebook's child privacy features. At the macro-level, we treat entire ad networks as our unit of analysis to see how GDPR-style rules on political advertising could affect the United States.

6.1 Macro-Level Tracking Trends

We believe our dataset to be one of the largest active measurements of web tracking to date conducted with a consumer web browser, and we collected it primarily as a basis to search for specific privacy and policy violations, rather than to conduct a macro-level census. However, it is useful for the wider literature to briefly note a few features of the overall set of 11.5 million page loads taken from 2.3 million sites.

One key area in which our measurement improves on prior studies is that our population is drawn from sites visited by real Chrome users. We also perform internal crawls of pages and limit ourselves to sites with no redirects and at least five internal pages. This means we are visiting fewer "junk" websites than other studies, many of which may have less tracking because they serve a placeholder function. We find 98% of pages expose users to an average of 21 third-party domains, 79% of sites expose users to an average of 12 cookie-setting third-party domains, 9% of sites have a web socket connection opened by an average of one third-party domain, and 45% of sites have DOM storage set by an average of 1.5 third-party domains. As additional restrictions are placed on third-party cookies by browsers, we believe websockets and DOM storage are areas to watch.

6.2 Macro-Level Features of Web Policies

Our primary motivations for collecting web polices at this scale are twofold: first, to provide a basis for searching for policy claims, and second, to expand the study of policies on the web beyond Privacy Policies alone. In pursuing these goals, we collected what we believe to be one of, if not the, largest corpus of current policies.[21] In this section, we present some macro-level findings of what this corpus tells us about polices on the web.

The first thing to note is the relative prevalence of different types of policies (see Table 1). We found Privacy Policies on 25% of sites, Terms of Service on 15%, Cookie Statements on 4%, and Ad Choices, GDPR, and CCPA Statements on fewer than 1%. We believe the reason for this is many sites have cookie, GDPR, and CCPA information in the Privacy Policy, which in many ways is preferable from a user standpoint. For example, the phrase "trade union membership" is fairly unique to the GDPR and we found that phrase in 2% of GDPR Statements and Privacy Policies, and fewer than 1% of other policies.

As has been studied extensively in the past, we find Privacy Policies are difficult to read (college graduate level on Flesch Reading Ease), and we extend prior findings to show that *all* types of

[20] Note we account for this and re-assign tasks until they are completed or a failure limit is reached

[21] Note another paper has recently collected one million historical Privacy Policies, itself a significant achievement, but these policies are not linked to tracking measurements as ours are.[1]

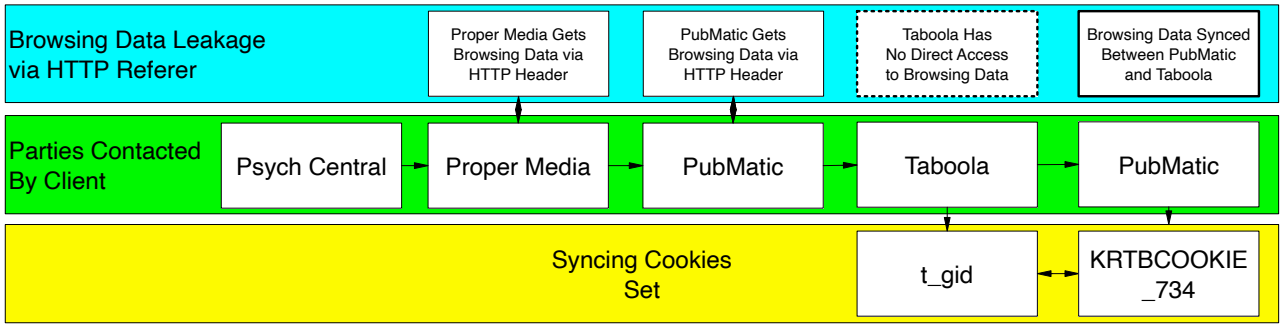

Figure 3: The Psych Central page for "Acute Stress Disorder Symptoms" exposes users to a chain of events by which Proper Media loads code from PubMatic who in turn loads an invisible iFrame from Taboola which sets a syncing cookie with PubMatic, allowing advertisements to be targeted on mental health.

policies are similarly difficult to understand, once again bringing doubts of their utility for most users, and reinforcing our belief their primary utility is in legal proceedings. Furthermore, we find Terms of Service to be the longest policies at 2,668 words.

Another factor we explored is "Pacifying Language", what we view as disingenuous statements meant to put the reader at ease before undesirable privacy practices are disclosed (or as one recent paper title put it, *"We Value Your Privacy ... Now Take Some Cookies"*[9]). We searched for policies containing the following terms: "we value", "we respect", "important to us", "help you", "we care", "committed to protecting", "cares about", and "transparency".[22] We found Privacy Policies (32%) and GDPR Statements (33%) were the most likely to have pacifying language and Terms of Service (13%) and cookie statements (10%) to have the least.

Finally, we examined the amount of tracking on pages with a given type of policy versus those without. We specifically wanted to know if the presence of a policy by itself was a potential signal the site had privacy issues. We find sites with Privacy Policies have 42% more third-party domains than those without, and sites with CCPA Statements have over four times as many.

6.3 Reverse Engineering a GDPR Violation

As noted above, relevant material evidence is needed to make a case in court. Our system is able to perform forensic measurements which save and hash files, record events at millisecond resolution, and facilitate reverse-engineering of specific events, such as cookies being set by a given party. For this section, we do a deep-dive on how a particular cookie gets set in violation of the GDPR.

In our analysis of third-party policies, we discovered that the company Taboola discloses they target ads based on "Personality - UK - Dealing with Stress - Emotional", a clear mental health topic. [23] While it will leave the E.U. shortly, the U.K. is in a "transition period" during which GDPR still applies, and we found it curious that this targeting segment exists as GDPR Article 29 states that "Processing of personal data...concerning health...shall be prohibited".

Given stress is a mental health topic, we used a haystack scan to find mental health websites Taboola could track, and thereby deduce a user's mental health status. We found a Taboola tracker on Psych Central, a website which claims to be the "Internet's largest and oldest independent mental health online resource"[24]. Psych Central is a top search result for many mental health conditions and touts endorsements by the New York Times and others. We then visited Psych Central and found a page for "Acute Stress Disorder Symptoms", which had a Taboola cookie when viewed from the United States, where it is completely legal to target ads based on medical conditions.

As noted above, unlike the US, targeting ads based on mental health is not legal where GDPR applies. Having located a potential GDPR violation, we instructed a measurement node located at a residential address in the U.K. to conduct a forensic scan of the Psych Central page for "Acute Stress Disorder Symptoms" to see if we could forensically document and reverse-engineer the process by which a Taboola cookie would be set on a mental health website.

On the basis of the forensic scan we captured requests initiated to 126 third-party domains, 6 third-party DOM storage entires, the content of 338 files loaded by the page, and 208 third-party cookies. One of these cookies, on the ".taboola.com" domain, is named "t_gid" and according to Taboola is a "unique User ID that allows Taboola to recommend specific advertisements and content to [a] user".[25] The value for "t_gid" was found to be a substring of the cookie "KRTBCOOKIE_734" set from the ".pubmatic.com" domain. According to PubMatic, this cookie allows them to "correlate our user IDs with those of our partner".[26] The overlap between cookies values, along with their stated purpose, indicated PubMatic and Taboola are collaborating to track mental health status.

Our forensic scan contains timestamps and file hashes for every event which occurred while loading the Psych Central page, all of which may be admitted to court. Figure 3 gives a visual overview of the following steps in the process:

[22]Note that while we collect policies in a variety of languages, we only searched for these terms in English, thus we can't capture the full extent without a larger vocabulary.

[23]Taboola is a member of the self-regulatory group Network Advertising Initiative (NAI), which requires members to disclose health targeting categories. It is important to note that Taboola is more forthcoming in providing targeting categories than other NAI members, who may also be engaged in similar practices but fail to reveal them.

[24]https://psychcentral.com/about/

[25]https://policies.taboola.com/en/cookie-policy/

[26]https://pubmatic.com/legal/platform-cookie-policy/

(1) Browser navigates to https://psychcentral.com/disorders/acute-stress-disorder-symptoms/
(2) Source code of Psych Central downloads Javascript from Proper Media
(3) Proper Media downloads file from PubMatic, receives HTTP referer revealing URL being visited
(4) PubMatic loads additional code, receives HTTP referer revealing URL being visited
(5) PubMatic creates invisible iFrame in which Taboola code is loaded, Taboola is unable to see URL being visited
(6) Taboola iFrame creates "t_gid" cookie
(7) Taboola iFrame redirects to PubMatic, passes value of "t_gid" as a URL paramater
(8) PubMatic sets syncing cookie named "KRTBCOOKIE_734" with the value of "t_gid" as a substring

Once the above sequence completes, a user's interest in "Acute Stress Disorder Symptoms" is exposed to Taboola, a company which specifically allows for user-level targeting against against stress, a mental health condition. In addition to this micro-level privacy violation, we can also surface several meso-level violations.

6.4 Finding Children's Privacy Violations

The 1998 Children's Online Privacy Protection Act (COPPA) is a US federal law which prohibits online services directed to children under 13 years old from collecting, using, or disclosing children's personal information. According to COPPA, collecting data includes encouraging a child to submit personal information online, enabling a child to make personal information publicly available, and passively tracking a child online. FTC guidance on COPPA states that sites subject to the law must "post a clear and comprehensive online Privacy Policy", and "provide direct notice to parents and obtain verifiable parental consent, with limited exceptions, before collecting personal information online from children".[27] Given these requirements, we sought to find any websites claiming adherence to COPPA and collecting user information prior to consent being given.

We focused on a specific aspect of children's online privacy: the use of Facebook code on child-directed websites. According to FTC guidance, social media code is allowed to collect data without parental consent *only* if the following conditions are met:

> ** a third-party operator only collects a persistent identifier and no other personal information;*
> ** the user affirmatively interacts with that third-party operator to trigger the collection; and*
> ** the third-party operator has previously conducted an age-screen of the user, indicating the user is not a child.*[28]

When a user loads a page, there may be no affirmative interaction with Facebook, and there is no generalizable way for Facebook to conduct an age-screen. For this reason Facebook provides "alternative code" that websites for children "are required to use for...Social Plugins and the Facebook SDK for JavaScript in the United States".[29] Specifically, sites for children using the JavaScript SDK must set the variable "kidDirectedSite" to "true" and those using the Like Button plug-in must set the "kid_directed_site" URL parameter to "true".

Starting with 2.3 million sites, we sought to find any that could be in violation of COPPA by using Facebook's Social Plugins without the appropriate flag. Our search criteria for including a site is as follows: the site must have the words "kid" or "child" in the title of the homepage, expose user information to a Facebook-owned domain, have a policy which affirmatively states COPPA compliance, and may be judged a child-directed site according to FTC guidance.[30] Our search space includes 1,098,751 sites tracked by Facebook, 32,908 sites with COPPA in a policy, and 10,202 sites with "kid" or "child" in the homepage title. However, the number of sites meeting all three criteria was 242. Once narrowed down to a manageable number, we were able to manually evaluate if policy text included an affirmative statement of compliance and if the site could be considered child-directed.

In regards to our manual classification of policy language, we included sites making statements that described or mentioned "complying" with COPPA, sites making semi-ambiguous statements such as "we do not *knowingly* collect PII from children under the age of 13"[31] (emphasis added), but excluded sites that claimed to not be child-directed if such an argument could feasibly be made. For example, we excluded a site for parents to evaluate toy purchases that claimed to be for users over 14, but included a site that claimed to not be child-directed in the Privacy Policy, but on the homepage stated "We have created this site with...preschoolers, and elementary school students in mind".[32] In regards to child-directed classification, the FTC guidance is fairly broad, including "the use of animated characters or child-oriented activities and incentives" and "age of models," and when in doubt, we took an inclusive approach. The manual process reduced the number of sites in question to only 107 sites.

Once target sites were identified, we conducted forensic scans from a US residential IP address. 106 sites successfully loaded (1 had an expired certificate), and of those, 42 were found to be using the Facebook JavaScript SDK. Only 2 of the 42 child-directed sites use the "kidDirectedSite" setting. Figure 4 shows a site which implemented the code correctly [33], and one which did not. Five sites use the "Like" button plugin (2 of which also used Javascript SDK, meaning three additional sites total), and none of these used the appropriate "kid_directed_site" URL parameter. In sum, 95% of sites targeted to children which claim COPPA compliance and include Facebook social code fail to properly implement Facebook's child protections.

We also found the first-party "_fbp" cookie was set by Facebook on 72 sites, the value of which was also transferred to the third-party domains "vipkid.com.cn" (on "http://lingobus.com") and "fullstory.com" (on "http://kiddom.co") as POST data, meaning they

[27] https://www.ftc.gov/tips-advice/business-center/guidance/complying-coppa-frequently-asked-questions-0
[28] https://www.ftc.gov/tips-advice/business-center/guidance/complying-coppa-frequently-asked-questions-0
[29] https://developers.facebook.com/docs/plugins/restrictions#child-directed
[30] https://www.ftc.gov/tips-advice/business-center/guidance/complying-coppa-frequently-asked-questions-0
[31] https://www.allfreekidscrafts.com/index.php/hct/privacy_policy
[32] https://kidcourses.com
[33] Note that in Javascript "!0" evaluates to "true".

Figure 4: The site on the left (Homer) uses Facebook's "kidDirectedSite" setting, the site on the left (Code Monkey) does not.

may also be collecting COPAA-protected data. This is despite the fact that Fullstory claims to allow "private customer data to be blocked at the source".[34]

While we only focused on Facebook, other COPPA violations likely exist well. The website AtlasMission states in their Privacy Policy that they are the "only organization collecting personal information regarding your child through this website"[35], yet they leak user data to 13 third-party domains. Other violations aside, by drawing 43 websites with a highly specific violation out of a corpus of 2.3 million sites, we demonstrate the power of our tool to find needles in the haystack. We also raise the question of who is responsible: the sites for potentially making a coding error, or Facebook, for not finding these issues before we did?

6.5 Potential Effects of GDPR-Style Political Ad Targeting Regulation in the US

While the above case studies focused on potential legal violations, we now shift focus to examine practices which *are entirely legal*, but perhaps should not be. While there are various forms of oversight of political spending in the United States, there are virtually no federal-level regulations which address targeted political advertising. This raises the possibility that *some* citizens may be shown advertisements tailored to their specific interests and attitudes, thereby denying *all* citizens equal access to political messaging. The second main affordance of our tool is to model how a proposed regulation would impact the web on a macro-level.

The lack of federal oversight and controversies such as Facebook's sharing of user data with Cambridge Analytica have led to calls for greater regulation of online political advertising.[36] One way to inform how future regulation could work is to apply model regulation to the extant US system. In this case, we utilize components of GDPR as a lens to highlight the potential effect of adopting similar legislation in the United States – specifically we seek to identify what current practices may be banned.

There are few avenues to determine exactly which companies engage in targeted political advertising, but the NAI has a code of conduct stating "segments" (meaning groups of users) used for targeted political advertising must be disclosed. Leveraging our corpus of third-party policies, we determined that 11 NAI member companies specify they allow targeting political ads. One company, Clickagy, claims to utilize "non-sensitive politically related segments," which they do not disclose, in potential violation of NAI requirements.

Having identified the targeting categories used by ten companies, we utilize Article 9 of the GDPR ("Processing of special categories of personal data") to identify five categories of prohibited data types pertinent to political advertising in the US: race, political opinions, religion, union membership, and sexual orientation. We add three topics not in the GDPR, but with particular salience to the US: personal wealth[37], viewpoints on gun control, and abortion. We determine if each company allows targeting in a given category by examining disclosed segments which are presented in Table 2. We likewise determine the total number of sites tracked by companies engaged in targeting on a given topic. The total number of sites represents not only where tracking data is harvested, but where ads could potentially be shown. For several categories, this is in excess of half a million sites, showing the impact, and likely monetary value, of using political tracking data on a macro-level.

In addition to total number of sites tracked, we also use several key terms based on advertising segment lists to find sites related to a given category to show how prevalent tracking is on certain types of sites. We use the following terms for each category:

- Race: *african american, asian american, chinese american*
- Political Views: *republican party, democratic party, green party, libertarian party*
- Religion: *catholic, christian, church, jewish, judaism, synagoge, islam, muslim, mosque, buddhist*
- Trade Unions: *(work or trade or teacher) and union, (excluding credit union)*
- Sex Orientation: *lgbt*,[38] *lesbian, gay, bisexual, transgender, queer, homosexual*
- Guns: *2nd amendment, second amendment, gun, firearm, rifle, pistol*
- Abortion: *prolife, pro-life, prochoice, pro-choice, abortion, planned parenthood, family planning, birth control*
- Wealth: *credit card, stock market, investment, bank, loan, debt*

[34] https://www.fullstory.com/responsibility/
[35] https://www.atlasmission.com/privacy-policy/
[36] https://thehill.com/blogs/congress-blog/politics/519933-its-time-for-congress-to-regulate-political-advertising-on

[37] While less prominent in the general election, income inequality was a major theme in Democratic primaries.
[38] Note the substring "lgbt" matches other variations such as "lgbtq".

	EU Excluded	Race	Political View	Religion	Trade Union	Sex Orient	Guns	Abortion	Wealth
Amobee	✓	✓	✓	✓		✓	✓	✓	✓
Choozle		✓	✓	✓		✓	✓	✓	✓
Foursquare	✓								
Lotame	✓		✓				✓		
Media Math	✓	✓	✓	✓	✓	✓	✓	✓	✓
Neustar			✓						
Taboola			✓			✓			
The Trade Desk		✓	✓	✓		✓	✓	✓	✓
Throtle	✓		✓						
Xandr		✓	✓				✓	✓	
Total Companies	5	5	9	4	1	5	6	5	4
% Topical Sites Tracked		38%	35%	19%	9%	15%	21%	44%	17%
Total Sites Tracked		523,101	556,567	480,939	322,386	436,568	524,896	523,101	421,848

Table 2: We found 10 companies disclosing political advertising segments, nine of which include targeting which violates the GDPR. Five companies specifically disallow political targeting in the EU, suggesting adopting similar rules could halt the practice elsewhere.

Note that the above terms sets are used primarily for illustrative purposes, and the related findings in Table 2, *% Topical Sites Tracked*, should be taken as an indication of what deeper study may reveal rather than a comprehensive evaluation. Indeed, one main benefit of macro-level analysis is to tease out areas that merit further inspection. For example, we found 44% of abortion-related sites were tracked by companies that target based on abortion viewpoints, and upon deeper inspection we found seven abortion providers with political tracking. [39]

The upshot of our findings are straight-forward: 5 companies are ready to comply with GDPR-based political advertising regulation *today*, the overall impact would be felt across over half a million sites, and visits to abortion clinic websites will no longer determine what political messages an individual receives. We imagine many Americans would welcome such a change.

7 DISCUSSION

Developing the system this paper is based on was a non-trivial task. However, since it has been up and running, being able to find highly-specific privacy violations across a huge swath of the web is incredibly exciting. The biggest thrill is being able to go from a hypothetical ("Does Taboola track any mental health websites?"), to an answer in mere seconds, and full forensic documentation in minutes. The biggest challenge in writing this paper was pulling ourselves out of many intriguing rabbit holes, some of which will spawn future papers.

We believe this type of open-ended exploration may be the key affordance our system brings to privacy regulation and litigation. While computer scientists have their own professional biases as to what makes a tracking violation interesting or "novel", lawyers ultimately know how to build a case that can stand up in court. In cases where a question is clearly defined, our system can surface the answer quickly. But we also allow for exploration, and those with a nose for a winnable lawsuit could surely sniff out points of action or follow a hunch.

[39] We are in the process of notifying them about this.

More than anything, our wish is for the companies putting users under the microscope to face the same uncomfortable scrutiny: we call this *regulatory surveillance*. One classic negative impact of surveillance is termed "chilling effects"; those afraid of being watched may hesitate to engage in certain activities such as exploring their sexual orientation or engaging in political action. Given we can collect and hash every piece of Javascript deployed over millions of sites on a rapid basis, it is easy for us to keep track of every line of code a company deploys to the web, and every change they make. Any time a new cookie is deployed, regulators and litigants could get an email or alert on their phone. We can only imagine the chilling effect on trackers if every new cookie resulted in thousands of lawyers around the world getting pinged.

Likewise, if a company were placed under a consent decree, it would be possible to monitor all coding changes they make and have them submit documentation and justification to authorities prior to deployment, or pay a fine for distributing privacy-violating code in violation of agreements with regulators. This may cause them to think twice. Indeed, the biggest privacy fine to date, $5B to Facebook, was based on violating a prior order. We believe proper application of our tool in the right hands could make serious inroads into the ultimate means of stopping tracking: making it unprofitable.

8 CONTRIBUTOR NOTE

Anokhy Desai contributed to legal review for evidence requirements and COPPA; Dev Patel contributed to literature review; Timothy Libert did remainder of work.

REFERENCES

[1] Amos, R., Acar, G., Lucherini, E., Kshirsagar, M., Narayanan, A., and Mayer, J. Privacy policies over time: Curation and analysis of a million-document dataset. *Preprint* (2020).

[2] Auxier, B., Rainie, L., Anderson, M., Perrin, A., Kumar, M., and Turner, E. Americans and privacy: Concerned, confused and feeling lack of control over their personal information. *Pew Research Center: Internet, Science & Tech (blog). November 15* (2019), 2019.

[3] Bhagyashree, R. The tor project on browser fingerprinting and how it is taking a stand against it, September 2019.

[4] Bratus, S., Lembree, A., and Shubina, A. Software on the witness stand: what should it take for us to trust it? In *International Conference on Trust and Trustworthy Computing* (2010), Springer, pp. 396–416.

[5] Casey, E. *Digital evidence and computer crime: Forensic science, computers, and the internet.* Academic press, 2011.

[6] Cranor, L. F. *Design and Evaluation of a Usable Icon and Tagline to Signal an Opt-Out of the Sale of Personal Information as Required by CCPA.* PhD thesis, University of Michigan, 2020.

[7] Cyphers, B. Sharpening our claws: Teaching privacy badger to fight more third-party trackers. *Eletronic Frontier Foundation* (2019).

[8] Dabrowski, A., Merzdovnik, G., Ullrich, J., Sendera, G., and Weippl, E. Measuring cookies and web privacy in a post-gdpr world. In *Passive and Active Measurement* (Cham, 2019), D. Choffnes and M. Barcellos, Eds., Springer International Publishing, pp. 258–270.

[9] Degeling, M., Utz, C., Lentzsch, C., Hosseini, H., Schaub, F., and Holz, T. We value your privacy... now take some cookies: Measuring the gdpr's impact on web privacy. *arXiv preprint arXiv:1808.05096* (2018).

[10] Efroni, Z., Metzger, J., Mischau, L., and Schirmbeck, M. Privacy icons: a risk-based approach to visualisation of data processing. *Eur. Data Prot. L. Rev. 5* (2019), 352.

[11] Emami-Naeini, P., Agarwal, Y., Cranor, L. F., and Hibshi, H. Ask the experts: What should be on an iot privacy and security label? *arXiv preprint arXiv:2002.04631* (2020).

[12] Englehardt, S., and Narayanan, A. Online tracking: A 1-million-site measurement and analysis. In *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security* (New York, NY, USA, 2016), CCS '16, Association for Computing Machinery, p. 1388–1401.

[13] Fiore, U., Castiglione, A., De Santis, A., and Palmieri, F. Countering browser fingerprinting techniques: Constructing a fake profile with google chrome. In *2014 17th International Conference on Network-Based Information Systems* (2014), pp. 355–360.

[14] Goldberg, R. Most americans continue to have privacy and security concerns, ntia survey finds. *National Telecommunications and Information Administration* (2018).

[15] Isaak, J., and Hanna, M. J. User data privacy: Facebook, cambridge analytica, and privacy protection. *Computer 51*, 8 (2018), 56–59.

[16] Johnson, G., Shriver, S., and Goldberg, S. Privacy & market concentration: Intended & unintended consequences of the gdpr. *Available at SSRN 3477686* (2020).

[17] Kelley, P. G., Bresee, J., Cranor, L. F., and Reeder, R. W. A "nutrition label" for privacy. In *Proceedings of the 5th Symposium on Usable Privacy and Security* (New York, NY, USA, 2009), SOUPS '09, Association for Computing Machinery.

[18] Leon, P., Ur, B., Shay, R., Wang, Y., Balebako, R., and Cranor, L. Why johnny can't opt out: A usability evaluation of tools to limit online behavioral advertising. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (New York, NY, USA, 2012), CHI '12, Association for Computing Machinery, p. 589–598.

[19] Libert, T. Exposing the invisible web: An analysis of third-party http requests on 1 million websites. *International Journal of Communication 9* (2015).

[20] Libert, T. An automated approach to auditing disclosure of third-party data collection in website privacy policies. In *Proceedings of the 2018 World Wide Web Conference* (Republic and Canton of Geneva, CHE, 2018), WWW '18, International World Wide Web Conferences Steering Committee, p. 207–216.

[21] Libert, T., and Binns, R. Good news for people who love bad news: Centralization, privacy, and transparency on us news sites. In *Proceedings of the 10th ACM Conference on Web Science* (New York, NY, USA, 2019), WebSci '19, Association for Computing Machinery, p. 155–164.

[22] Libert, T., Graves, L., and Nielsen, R. K. Changes in third-party content on european news websites after gdpr. *Whitepaper* (2018).

[23] Malloy, M., Koller, J., and Cahn, A. Graphing crumbling cookies. *AdKDD '19* (2019).

[24] Marotta, V., Abhishek, V., and Acquisti, A. Online tracking and publishers' revenues: An empirical analysis. In *Workshop on the Economics of Information Security* (2019).

[25] Marthews, A., and Tucker, C. E. Government surveillance and internet search behavior. *Available at SSRN 2412564* (2017).

[26] Mathur, A., Vitak, J., Narayanan, A., and Chetty, M. Characterizing the use of browser-based blocking extensions to prevent online tracking. In *Proceedings of the Fourteenth USENIX Conference on Usable Privacy and Security* (USA, 2018), SOUPS '18, USENIX Association, p. 103–116.

[27] McDonald, A. M., and Cranor, L. F. The cost of reading privacy policies. *Isjlp 4* (2008), 543.

[28] Merzdovnik, G., Huber, M., Buhov, D., Nikiforakis, N., Neuner, S., Schmiedecker, M., and Weippl, E. Block me if you can: A large-scale study of tracker-blocking tools. In *2017 IEEE European Symposium on Security and Privacy (EuroS P)* (2017), pp. 319–333.

[29] Pochat, V. L., Van Goethem, T., Tajalizadehkhoob, S., Korczyński, M., and Joosen, W. Tranco: A research-oriented top sites ranking hardened against manipulation. *arXiv preprint arXiv:1806.01156* (2018).

[30] Roesner, F., Kohno, T., and Wetherall, D. Detecting and defending against third-party tracking on the web. In *9th USENIX Symposium on Networked Systems Design and Implementation (NSDI 12)* (San Jose, CA, Apr. 2012), USENIX Association, pp. 155–168.

[31] Shen, Y., and Vervier, P.-A. Iot security and privacy labels. In *Privacy Technologies and Policy* (Cham, 2019), M. Naldi, G. F. Italiano, K. Rannenberg, M. Medina, and A. Bourka, Eds., Springer International Publishing, pp. 136–147.

[32] Sørensen, J., and Kosta, S. Before and after gdpr: The changes in third party presence at public and private european websites. In *The World Wide Web Conference* (New York, NY, USA, 2019), WWW '19, Association for Computing Machinery, p. 1590–1600.

[33] Stoycheff, E. Under surveillance: Examining facebook's spiral of silence effects in the wake of nsa internet monitoring. *Journalism & Mass Communication Quarterly 93*, 2 (2016), 296–311.

[34] Turow, J. *The Aisles Have Eyes: How Retailers Track Your Shopping, Strip Your Privacy, and Define Your Power.* Yale University Press (Ignition), 2017.

[35] Ur, B., Leon, P. G., Cranor, L. F., Shay, R., and Wang, Y. Smart, useful, scary, creepy: perceptions of online behavioral advertising. In *proceedings of the eighth symposium on usable privacy and security* (2012), pp. 1–15.

[36] Urban, T., Degeling, M., Holz, T., and Pohlmann, N. Beyond the front page:measuring third party dynamics in the field. In *Proceedings of The Web Conference 2020* (New York, NY, USA, 2020), WWW '20, Association for Computing Machinery, p. 1275–1286.

[37] Van Kleek, M., Liccardi, I., Binns, R., Zhao, J., Weitzner, D. J., and Shadbolt, N. Better the devil you know: Exposing the data sharing practices of smartphone apps. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems* (New York, NY, USA, 2017), CHI '17, Association for Computing Machinery, p. 5208–5220.

[38] Zaeem, R. N., German, R. L., and Barber, K. S. Privacycheck: Automatic summarization of privacy policies using data mining. *ACM Trans. Internet Technol. 18*, 4 (Aug. 2018).

[39] Zimmeck, S., Story, P., Smullen, D., Ravichander, A., Wang, Z., Reidenberg, J., Russell, N. C., and Sadeh, N. Maps: Scaling privacy compliance analysis to a million apps. *Proceedings on Privacy Enhancing Technologies 2019*, 3 (01 Jul. 2019), 66 – 86.

[40] Zimmeck, S., Wang, Z., Zou, L., Iyengar, R., Liu, B., Schaub, F., Wilson, S., Sadeh, N. M., Bellovin, S. M., and Reidenberg, J. R. Automated analysis of privacy requirements for mobile apps. In *NDSS* (2017).