To appear in "Transparent data mining for Big and Small Data". Editors:
Tania Cerquitelli, Daniele Quercia, Frank Pasquale. Springer, 2017.

The Princeton Web Transparency and Accountability Project

Arvind Narayanan and Dillon Reisman

Abstract When you browse the web, hidden "third parties" collect a large amount of data about your behavior. This data feeds algorithms to target ads to you, tailor your news recommendations, and sometimes vary prices of online products. The network of trackers comprises hundreds of entities, but consumers have little awareness of its pervasiveness and sophistication. This chapter discusses the findings and experiences of the Princeton Web Transparency Project (https://webtap.princeton.edu/), which continually monitors the web to uncover what user data companies collect, how they collect it, and what they do with it. We do this via a largely automated monthly "census" of the top 1 million websites, in effect "tracking the trackers". Our tools and findings have proven useful to regulators and investigatory journalists, and have led to greater public awareness, the cessation of some privacy-infringing practices, and the creation of new consumer privacy tools. But the work raises many new questions. For example, should we hold websites accountable for the privacy breaches caused by third parties? The chapter concludes with a discussion of such tricky issues and makes recommendations for public policy and regulation of privacy.

1 Introduction

In 1966, Marvin Minsky, a pioneer of artificial intelligence, hired a freshman undergraduate for a summer to solve 'the vision problem', which was to connect a TV camera to a computer, and get the machine to describe what it sees [1]. Today this anecdote is amusing, but only because we understand with the benefit of hindsight that many, or even most things that are easy for people are extraordinarily hard for

Arvind Narayanan
Princeton University, Princeton NJ, e-mail: arvindn@cs.princeton.edu

Dillon Reisman
Princeton University, Princeton NJ, e-mail: dreisman@princeton.edu

computers. So the research field of AI barked up a few wrong trees, and even had a couple of so-called "AI winters" before finally finding its footing.

Meanwhile, automated decision-making had long been recognized by industry as being commercially valuable, with applications ranging from medical diagnosis to evaluating loan applications. This field too had some missteps. In the 1980s, billions of dollars were invested in building "expert systems", which involved laboriously creating databases of facts and inference rules from which machines could supposedly learn to reason like experts. An expert system for medical diagnosis, for example, would rely on physicians to codify their decision-making into something resembling a very large decision tree. To make a diagnosis, facts and observations about the patient would be fed into this set of rules. While such expert systems were somewhat useful, they ultimately they did not live up to their promise [2].

Instead, what has enabled AI to make striking and sustained progress is "machine learning". Rather than represent human knowledge through symbolic techniques, as is done in expert systems, machine learning works by mining human data through statistical means — data that is now largely available thanks to the "Big Data" revolution. Machine learning has become a Silicon Valley mantra and has been embraced in all sorts of applications. Most famously, machine learning is a big part of online personalized advertising — as data scientist Jeff Hammerbacher said, "The best minds of our generation are thinking about how to make people click on ads."[3]

But machine learning has also made its way into more important applications, like medical diagnosis and the determination of creditworthiness. It's even being employed in "predictive policing" and the prediction of criminal recidivism, where it has life-or-death consequences [4]. Machine learning has also proved extremely adept at the computer vision problems that Minsky was interested in fifty years ago, as well as related domains such as natural-language processing.

So machine intelligence and automated decision-making today increasingly rely on machine learning and big data. This brings great benefits ranging from movie recommendations on Netflix to self-driving cars. But it has three worrying consequences. The first is privacy. Many machine learning systems feed on people's personal data — either data about them or data created by them. It can be hard to anticipate when and how a piece of personal data will be useful to make decisions, which has led to a culture of "collect data first, ask questions later." This culture has spurred the development of an online surveillance infrastructure that tracks, stores, and profiles everything we do online.

The second consequence to society is that the outputs of machine learning reflect human biases and prejudices. One might have naively assumed that AI and machines in the course of automated decision-making would somehow be mathematically pure and perfect, and would make unbiased decisions. Instead, we're finding that because the data used to train machine learning models comes from humans, machines essentially inherit our biases and prejudices [5].

The third concern over our use of machine learning is what you might call the inscrutability of AI. It may have once been correct to think of automated decision-making systems as some sort of decision tree, or applying a set of rules. That's simply not how these systems work anymore. When a machine learning system is

trained on a corpus of billions of data points to make medical diagnoses, or to serve ads online, it is impossible to express to a patient or a user, or even to the creator of the system, the reason why the machine decided as it did. When we put these complex, inscrutable decision-making systems in a position of power over people, the result can be Kafkaesque [6].

These three concerns have led to much-needed public and scholarly debate. In addition, the inscrutability of these systems has necessitated a new kind of empirical research that can peek into the black boxes of algorithmic systems and figure out what's going on. This new research field, which we call "data and algorithmic transparency", seeks to address questions about the data collection and use that happens around us everyday, questions such as:

- Are my smart devices in my home surreptitiously recording audio?
- Does my web search history allow inferring intimate details, even ones I've not explicitly searched for?
- Is the algorithm that decides my loan applications biased?
- Do I see different prices online based on my browsing and purchase history?
- Are there dangerous instabilities or feedback loops in algorithmic systems ranging from finance to road traffic prediction?

A combination of skills from a variety of areas of computer science is called for if we are going to learn what is inside the black boxes. We need to build systems to support large-scale and automated measurements, and modify devices to record and reverse engineer network traffic. We need expertise to simulate, model, and probe decision-making systems. And we need to reach out to the public to collect data on the behavior of these algorithmic systems in the wild.

But more than just computer scientists, this effort is bringing together a new interdisciplinary community of empirical researchers, journalists, and ethical scholars, with new conferences and workshops focusing on transparency research, such as the workshop on Data and Algorithmic Transparency (http://datworkshop.org/). In an ideal world, one would imagine that companies involved in building data-driven algorithms would be perfectly forthcoming about what personal data they're using and how these systems work behind the scenes, but unfortunately that is not the world we live in today. Even if companies and governments were forthcoming, analyzing and understanding the societal impact of these systems will always require empirical research and scholarly debate.

2 The Princeton Web Transparency and Accountability Project

We started the "Princeton Web Transparency and Accountability Project" (WebTAP) to focus our study on the web, a rich source of data that feeds into decision-making systems, and a prominent arena where their effects can be seen. A major output of WebTAP is the "Princeton Web Census" which is an automated study of privacy across 1 million websites which we conduct every month.

Our work so far has focused on monitoring and reverse-engineering web tracking. Given the inadequacy of laws and rules that govern web tracking, we believe that external oversight of the online tracking ecosystem is sorely needed. We're interested in questions like, "Which companies track users? What technologies are they using? What user data is being collected? How is that data being shared and used?"

2.1 The perils of web tracking

What are the potential harms of web tracking? The first is simply that the erosion of privacy affects our intellectual freedom. Research shows when people know they're being tracked and surveilled, they change their behavior [7]. Many of today's civil liberties — say, marriage equality — were stigmatized only a few decades ago. The reason it became possible to discuss such issues and try to change our norms and rules is because we had the freedom to talk to each other privately and to find like-minded people. As we move to a digital world, is ever-present tracking hindering those abilities and freedoms?

A second worrisome effect of web tracking is the personalization and potential discrimination that results from the use of our personal data in these algorithmic systems. Some online retailers have experimented with price discrimination, showing different prices to different visitors based on personal factors [8]. Without oversight, we lose the ability to limit practices that would be censured in the offline world.

There are many other domains which might be impacted through algorithmic personalization. Privacy scholar Ryan Calo argues that personalized digital advertising can be construed as a form of market manipulation [9]. There are also impacts in the political sphere, with consequences for the health of democracy. What happens when political campaigns start to use personal data in order to microtarget the messages that they send to voters, to tell slightly different stories to different people online through targeted advertisements?

Finally, the lack of transparency in online tracking is problematic in and of itself. There's no public input into the decision-making that happens in these systems, leading to unaccountable and opaque processes that have real consequences. We need to close that transparency gap. This is a large part of what motivates the Princeton WebTAP project.

2.2 How web tracking works

In WebTAP, we mostly study "third party online tracking". When you go to nytimes.com, the New York Times knows you've visited and which article you're reading — in this case, the New York Times is a "first party". Because you choose to visit a first party, we are not particularly concerned about what the first party

knows from your visit. Third party online tracking, however, happens when entities other than the one you're currently visiting compile profiles of your browsing history without your consent or knowledge [10]. While most third parties are invisible, visible page elements such as Facebook Like buttons, embedded Twitter feeds, and a variety of other commercial widgets are also modes of third party tracking. One study a few years ago showed that, on the average top 50 website, there are 64 independent tracking mechanisms [11]! That is consistent with our own findings — in fact, that number has only grown over time [12].

Web cookies are the most widely used mechanism for tracking on the web by first and third parties, a fact many users are already familiar with. What is less well-known is that, increasingly, websites and trackers are turning to techniques like browser fingerprinting — techniques that are sneakier, harder for users to protect themselves from, and which work without necessarily leaving any trace on your computer. The *Beauty and the Beast* project (https://amiunique.org/) and the older *Panopticlick* project (https://panopticlick.eff.org/) offer a demonstration of fingerprinting using a variety of attributes from your own web browser, such as the type and version number of the browser, the list of fonts that you have installed, the list of browser plugins you have installed, and more [13, 14]. Third parties can use such fingerprints to uniquely identify your device, no cookies required.

You might think that this tracking is anonymous, since your real name is not attached to it. The online advertising industry has repeatedly sought to reassure consumers this way. But this is a false promise. Many third parties do know your real identity. For example, when Facebook acts as a third party tracker they can know your identity as long as you've created a Facebook account and are logged in — and perhaps even if you aren't logged in [15]. Third parties with whom you don't have an account have many ways of inferring a user's real identity as well. Sometimes all that is needed are bugs and poor web development, resulting in personal identifiers "leaking" from first parties to third parties [16, 17]. It is also possible for a tracker to de-anonymize a user by algorithmically exploiting the statistical similarity between their browsing history and their social media profile, as demonstrated in a recent collaboration between Stanford researchers and WebTAP [18].

Even ignoring all this, web tracking is not anonymous. "Anonymous" means that an individual's activities (say, visits to different websites) cannot be linked together, but such linking is the entire point of third-party web tracking. "Pseudonymous" means that those activities can be linked together, even if the real-world identity behind them is unknown. As Barocas and Nissenbaum have argued, most potentially undesirable effects of tracking happen even if you're being tracked under a pseudonymous identity instead of your real identity[19]. The potential biases and discrimination that we discussed — targeted political messaging, price discrimination, market manipulation — can still happen online, as long as advertisers have some way to communicate to you.

2.3 The Open Web

Third party tracking is rampant in part because the web is built on open technology, and so the barriers to entry are low. For example, once there is one third party on a page, that third party has the ability to turn around and invite any number of other third parties to the first party webpage. Web technology standards and specifications are the primary regulator of privacy practices, and these tend to be permissive. As we'll see in Section 4.1, new HTML features introduced by standards bodies have repeatedly been repurposed for infringing privacy in devious and inventive ways. In contrast, in a closed platform such as Apple's app store, Apple exercises significant control over what behaviors are acceptable [20].

But by the same token, the web's openness is also good news for users and browser privacy tools. The browser ultimately acts on behalf of the user, and gives you — through extensions — an extraordinary degree of control over its behavior. This enables powerful tracker-blocking extensions to exist, which we'll revisit in Section 4.4. It also allows extensions to customize web pages in various other interesting ways ranging from making colors easier on the eyes to blocking ads.

A recent WebTAP paper demonstrates how powerful this capability can be. We set out to analyze the so-called ad-blocking wars and to predict how ad-blocking technology might evolve [21]. Web publishers and other news sites are unhappy about the increasing popularity of ad-blockers among users. One response has been the deployment of ad-blocker blockers, to prevent users from viewing content unless the ad blocker is disabled. We argue that this strategy could not succeed in the long run — a determined user will always be able to block ads. Ad blocking extensions can modify the browser in powerful ways, including hiding their own existence from the prying JavaScript code deployed by websites. Ad blockers will likely ultimately succeed at this because browser extension code executes at a higher "privilege level" than website code.

Other websites, most notably Facebook, are trying to make their ads indistinguishable from regular posts, thus making it harder for ad-blockers to block ads without also blocking real user content. This is again a dubious strategy in the long run. Due to the enforcement of deceptive advertising rules by the US Federal Trade Commission, human users have to be able to tell ads and regular content apart. Ad industry self-regulatory programs such as AdChoices also have the same effect. We created a proof-of-concept extension to show that, if humans are able to distinguish ads and content, then automated methods could also be able to distinguish them, by making use of the same signals that humans would be looking at, such as the text "Sponsored" accompanying an advertisement [22, 23].

The broader lesson is that open technologies shift the balance of power to the technologically savvy. Another consequence of this principle is that the web's openness is good news for us as researchers. It makes the technical problem of automated oversight through web privacy measurement much easier. Let us give you a small but surprising example to illustrate this point.

A popular service provider on the web, called Optimizely, helps websites do "A/B testing." It is the practice of experimenting with different variations of a part of the

website to evaluate how positively users respond to the variations. For instance, Optimizely allows publishers like New York Times to test two different versions of a headline by showing one version to 50% of visitors, and another version to the other 50%, so that they can evaluate which headline more users clicked on. Because of the open nature of the web, the code Optimizely uses to implement the experiments is actually exposed on an invisible part of the webpage — any visitor on the page who knows how to peek behind the scenes of the web browser could see every experiment that the website was doing on its users!

So we did a little digging. As part of the WebTAP project, we visited a number of different sites, and grabbed all of the Optimizely code and experimental data from sites that used Optimizely. You can find the full details of our study in our blog post, but there were some interesting instances of A/B testing that stood out [24]. Many news publishers experiment with the headlines they feature, in a questionable way. For example, a headline such as "Turkey's Prime Minister quits in Rift with President" might appear to a different user as "Premier to Quit Amid Turkey's Authoritarian Turn." You can see clearly that the implication of the headline is different in the two cases. We also found that the website of a popular fitness tracker targets users that originate from a small list of hard-coded IP addresses labeled "IP addresses spending more than $1000." While amusing, this can also be seen as somewhat disturbing.

3 WebTAP's main engine: OpenWPM

Back in 2013, when we started WebTAP, we found that there had been over twenty studies that had used automated web browsers for studying some aspect of privacy or security, or online data-driven bias and discrimination. We found that many of those studies had devoted a lot of time to engineering similar solutions, encountering the same problems and obstacles. We were motivated to solve this problem and share it with the community, which led to our development of OpenWPM — Open Web Privacy Measurement.[1] It builds on ideas and techniques from prior projects, especially FourthParty [10] and FPDetective [25]. Today it is a mature open source project that has a number of users and researchers using it for a variety of studies on online tracking.

3.1 Problems solved by OpenWPM

OpenWPM solves a number of problems that researchers face when they want to do web privacy measurements. First, you want your automated platform to behave like a realistic user. Many researchers have used a stripped-down browser, such as

[1] OpenWPM is available for download at https://github.com/citp/OpenWPM

PhantomJS, that does not fully replicate the human browsing experience [26]. While this might be okay for some experiments, it won't reproduce what happens when a real user browses the web. To solve this problem the OpenWPM platform uses a full version of the popular Firefox web browser. [2]

OpenWPM also allows simulating users with different demographics or interest profiles by building up a specified history of activity in the automated browser. For example, before beginning an experiment a researcher can have OpenWPM visit sites that are popular among men or among women, depending on the profile they are trying to mimic.

The second problem that OpenWPM solves is that it collects all data that might be relevant to privacy and security concerns from the automated web browser. This includes all of the network requests the automated browser makes, including cookies, but also information that's relevant to browser fingerprinting. OpenWPM has the ability to log every single JavaScript call to a browser JavaScript API that is made by a script on the page, and the identity of the third party or first party that made the call. We have found this feature very useful in our own experiments.

This set of information is collected from three different vantage points. A browser extension collects the JavaScript calls as they are made in the browser. A "man in the middle" proxy intercepts network requests made between the browser and the website it visits. Lastly, OpenWPM collects any information that is stored on the computer's disk, such as flash cookies. OpenWPM unifies these views of the data and stores them in a single database that provides an easy interface for analysis.

OpenWPM also makes it easier for researchers when it comes time for data analysis. We provide tools that can aid researchers in answering simple questions like, "What are all of the third parties that were ever seen on a particular first party site?" More complex analyses can be easily built off of those building blocks as well.

The stability of the tools that are used for web measurement has also been a problem for past researchers. OpenWPM uses a tool called Selenium to automate the actions of the web browser, as done in many prior web privacy measurement experiments [27]. While Selenium makes issuing commands to Firefox very simple, it was meant for testing single websites. In our experiments we visit as many as one million websites every month. Selenium, we've discovered, was not made to scale to that many websites, and without "babysitting" by the researcher, will often crash in the middle of an experiment. OpenWPM solves the stability issue by recovering from crashes in Selenium while preserving the state of the browser.

3.2 OpenWPM's advanced features

We're also working on a goal we call "one-click reproducibility," which would give researchers the ability to reproduce the results of experiments done by others. OpenWPM issues commands that could be stored in a standard format and replayed later.

[2] It is possible to use another browser, such as Chrome, in OpenWPM, but in our experiments so far we've used Firefox.

A package containing these commands along with analysis scripts used to evaluate the experiments and its results could be shared to enable collaboration among researchers and verification of experimental results. One possible use-case for one-click reproducibility can be found in the world of regulation — if a regulator conducts a study involving a complex series of commands across a set of websites, then they may want the ability to share the exact tools and procedure they used with other parties concerned with the regulation. Beyond specific use-cases, we hope that this will help scientific replication of web privacy studies.

OpenWPM also has a number of advanced features that are specifically useful for privacy studies. OpenWPM can automatically detect which cookies are unique tracking cookies and hence privacy-relevant (as opposed to cookies setting website language, or timezone). OpenWPM also has a limited ability to automatically login to websites using "federated login mechanisms". Federated login mechanisms, like Facebook Connect or Google's Single Sign-On (SSO), are tools that, when added to a website, allow users of that website to use their Facebook or Google account to sign-in. Since many websites leverage federated login mechanisms, our ability to use it to automatically login can be quite useful since certain privacy impacting behaviors are only triggered when a user is logged in.

OpenWPM also has a limited ability to extract content from webpages. For example, if a researcher wants to measure how search results vary between users, there is a way for them to specify how to extract the search results from a webpage. While the process is not yet entirely automated, it only requires a few lines of code. OpenWPM stores such extracted content for later analysis.

4 WebTAP's Findings

Frequently, we've found, a privacy researcher publishes a study that finds a questionable privacy-infringing practice. This puts public scrutiny on the third parties or websites concerned, leading to a temporary cessation of the practice, only for the episode to be forgotten a month later, at which time the privacy-infringing behavior creeps back into use.

WebTAP avoids these problems of "one-off" privacy studies because our studies are *longitudinal*. Our analyses are automated, which allows us to continually monitoring questionable practices on the web and keep pressure on trackers to avoid those practices.

4.1 Detecting and measuring novel methods of fingerprinting

Through our crawls of the web we've found that third-party tracking technologies evolve rapidly. In particular, we've tracked the evolution of browser fingerprinting techniques. Most or all of the information provided by your browser to websites

and third parties can be used to fingerprint your device and to track you. As long as a piece of information is at least somewhat stable over time, it can be leveraged in a fingerprint. This can include things as innocuous as the size of your browser window, the version number of your browser, or even the set of fonts installed on your computer.

Imagine that an online tracker derives a fingerprint for you that consists of, say, 43 different features. After the tracker first sees you, however, you install two new fonts on your computer. When the tracker sees you again, you'll have a fingerprint with 45 features, of which 43 will match the previous observation. Much like observations of actual fingerprints, two such observations that differ slightly can still be linked through statistical means. In the industry, these fingerprints are often called "statistical IDs."

In a 2014 study, we collaborated with researchers at KU Leuven to study a technique called "canvas fingerprinting" [28]. The Canvas API, recently added to browsers in the HTML5 standard, gives scripts a simpler interface for drawing images on webpages. Scripts can also use the canvas to draw an image that is invisible to the user. Once the image is drawn on the webpage, the script can read the image data back pixel by pixel. It turns out that the precise pixel-wise representation of the image, such as the one seen in Figure 1, will vary between different devices based on unique features of the rendering software that is used for displaying images on the screen. The Canvas API is one of the sneakier ways that a seemingly benign interface provided by the browser can contribute to a unique fingerprint.

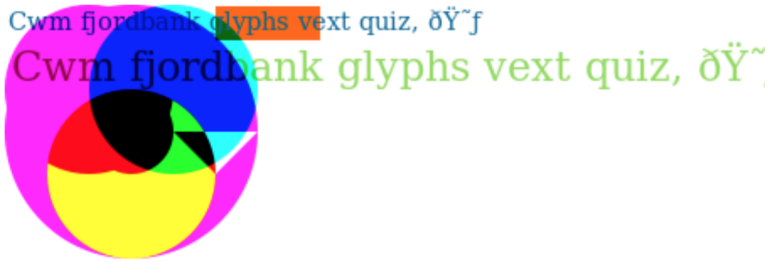

Fig. 1 The image, invisible to the user, that is drawn on the web page by one of the canvas fingerprinting scripts that we detected. The image is converted to a string of data, constituting the fingerprint.

Canvas fingerprinting had been first proposed in a security paper in 2012 [29]. Some time later a developer implemented an open-source library that included canvas fingerprinting, and a few obscure sites experimented with it [30]. But by the time of our study in early 2014, a mere year and a half after it had first been proposed, several third parties, including a major online third party called AddThis, had employed canvas fingerprinting — of the top 100,000 websites included in our study, over 5% had third party scripts that employed the technique. This is one example

of how obscure tracking techniques can quickly go from academic curiosities to becoming mainstream on the web. This discovery was only possible through the use of automated measurement. This quickly led to improvements in browser privacy tools as well as a public and press backlash that led AddThis to drop the technique.

This study was done in 2014, and didn't use OpenWPM, as it wasn't mature yet. In 2016, we began WebTAP's monthly 1-million-site measurements using OpenWPM. We've found that virtually every HTML5 API is being abused to fingerprint devices. This includes the AudioContext API for audio processing, the Battery Status API [31], and the WebRTC API for peer-to-peer real-time communication. We were able to catch these techniques early in their lifecycles, when they were deployed on relatively small numbers of sites [32]. Our ongoing findings have led to debate and discussion in the web standards community on how to design these APIs in a way that resists the ability to utilize them for fingerprinting [33].

4.2 The "collateral damage" of web tracking

WebTAP has shed light on how the online tracking infrastructure built by web companies for commercial purposes can be repurposed for government surveillance. The Snowden leaks revealed that the NSA has in fact been reading tracking cookies sent over networks to enable their own user tracking [34]. We wanted to study and quantify just how effective this could be.

The answer isn't obvious. First, the technique might not work because any given tracker might appear on only a small fraction of web pages — it is unclear to what extent the NSA or another surveillance agency might be able to put together a complete picture of any given person's web browsing traffic using just these cookies. Second, cookies are pseudonymous, as we discussed earlier. Even though Facebook, for example, might know the real-world identity associated with each cookie, it will not necessarily send that information across the network where the NSA can read it. There are various other complications: for example, it is not clear what portion of a typical user's traffic might pass through a vantage point on the internet that the NSA can observe.

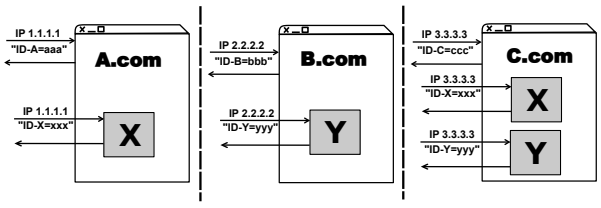

Fig. 2 How cookie linking works: An eavesdropper observes identifiers ID-X and ID-Y on two separate page loads. On a third page load, both identifiers are seen simultaneously, allowing the eavesdropper to associate ID-X and ID-Y as belonging to the same user.

Using OpenWPM, we simulated a typical user browsing the web and analyzed which trackers were embedded on which websites [35]. Even if a tracker is not embedded on a significant number of websites, if two different trackers are embedded on the same page, an eavesdropper on the network can infer that the two tracker's distinct IDs belong to the same user. With enough trackers embedded on enough websites, it is possible to transitively link all of the tracking IDs for a single user using the procedure illustrated in Figure 2.

Using this notion of transitive cookie linking, as well as using geographical measurements of where websites and their servers are located, we were able to show that an eavesdropper like the NSA will be able to reconstruct 60-70% of a user's browsing history. Since a number of sites have inadequate use of encryption on their websites, we were also able to show that such an adversary will very likely be able to attach real-world identities to these browsing histories for typical users. If a website fails to encrypt a page containing personally identifiable information, like a name or email address as seen on the New York Times website in Figure 3, we found that it was likely that eventually a real identity would be leaked across the network along with the tracking cookies.

Fig. 3 The email address of a logged-in New York Times user is displayed on the top banner of the website (highlighted in red). Because the New York Times does not deliver their webpage over an encrypted connection (as of September 2016), an eavesdropper can use the email address to assign a real-world identity to a pseudonymous web browser.

This finding highlights the fact that concerns about online tracking go beyond its commercial applications, and have consequences for civil society. It also underscores the importance of deploying encryption on the web — HTTPS is not just for protecting credit card numbers and other security-sensitive information, but also for protecting the privacy of the trails we leave online, which might otherwise be exploited by any number of actors.

Unfortunately, in the same theme of "collateral damage," our research has also shown that third parties on the web impede the adoption of HTTPS [32]. Our measurements reveal that many players in the online tracking ecosystem do not provide an encrypted version of their services. The web's security policies dictate, with good reason, that for a web page to deploy HTTPS, all the third-party scripts on the page also be served encrypted. About a quarter of unencrypted sites today would face problems transitioning to HTTPS due to third parties.

4.3 The economic forces behind tracking

WebTAP's findings are in line with the intuition that the need for ad revenue, especially among news publishers, is a big driver behind the prevalence of tracking [32]. Breaking down the number of trackers by website category, available in Figure 4, we see that news sites have the most embedded trackers. Interestingly, adult sites seem to have the least. Perhaps adult websites are more concerned with user privacy, or perhaps fewer advertisers are interested in advertising on adult sites; we cannot know for sure from the data alone.

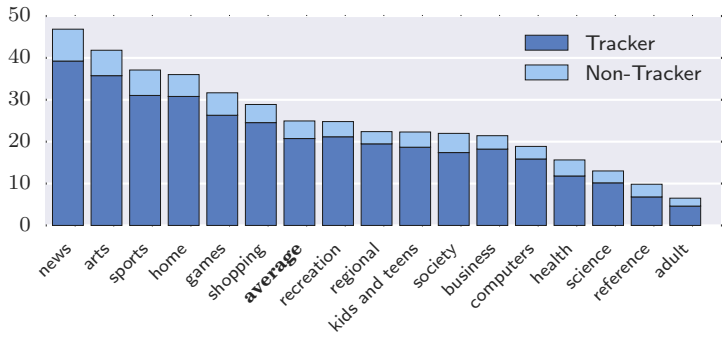

Fig. 4 The prevalence of third-parties by Alexa site category, split between tracking and non-tracking third parties.

We've also found that the web ecosystem has experienced some consolidation among tracking companies. There is certainly a long tail of online tracking — in our study of one million websites, we found that there are over 80,000 third parties that are potentially in the business of tracking. Most of these, however, are only found on a small number of first-party sites. In fact, there are relatively few third parties, around 120, that are found on more than 1% of sites. And a mere six companies have trackers on more than 10% of websites, as shown in Figure 5 [32].

Arguably, this is good for privacy, because the more popular third parties are larger companies like Google or Facebook that are more consumer-facing. When such companies get caught in a privacy misstep, it is more likely that the resulting negative public reaction will have consequences for them. These bigger players also get more scrutiny from regulatory agencies like the Federal Trade Commission and are, perhaps, more willing to police themselves. On the other hand, there's also an argument that the large concentrations of online tracking data that result from economic consolidation are bad for privacy.

Even ignoring economic consolidation, there are several ways in which different tracking databases get linked to or merged with each other, and these are unambiguously bad for privacy. Because of a process called "cookie syncing," individual

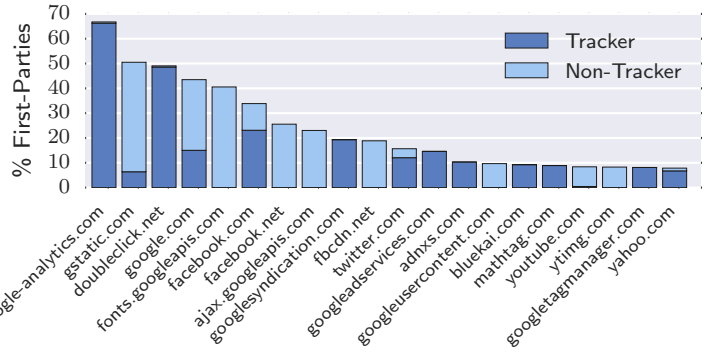

Fig. 5 The most prevalent third-parties by domain, split between when the third party was identified as tracking vs. non-tracking.

trackers gain a greater ability to see more of your online behavior. Cookie syncing allows trackers to link their identifying cookies to other companies' cookies, giving a tracker's customers greater insights into a user's activity from more sources. In our census, we found that 85 of the top 100 most common third parties sync their cookies with at least one other party. Google was the most prolific cookie-syncer — there are *118* other third parties with which it shares cookies or which share cookies with it.

Researchers and journalists have documented several other trends towards consolidation of tracking databases. First, the *internal* privacy walls at tech companies are coming down [36, 37, 38]. "Cross-device tracking" links your different devices with each other — sometimes by invasive methods such as playing an ultrasound audio code on a web page and listening to it from your smartphone [39, 40]. "Onboarding" by firms including Facebook links your (physical) shopping records with your online activity [41]. The euphemistic "header enrichment" by Internet Service Providers adds the equivalent of a cookie to your web or mobile traffic — a single cookie for all your activities rather than a tracker-specific identifier [42].

4.4 The impact of web privacy measurement

We've found repeatedly that merely measuring online tracking has a positive impact on web privacy, due to increased public awareness and companies' desire to avoid the spotlight on privacy issues. This was a surprise for us — traditionally in computer security research, measurement is at best a first step to try and understand the scope of the problem before you start the process of devising solutions.

Measurement seems to mitigate the "information asymmetry" that exists in online tracking. That is, web trackers know a lot more about the technologies and the

type of data that's being collected than consumers do. In fact, publishers are often in the dark about the extent of third party tracking on their own websites. In response to our study, many of the websites where canvas fingerprinting had been found responded to press inquiries to say that they were entirely unaware of the practice. Measurement seems to fix this information asymmetry for both users and website owners alike, making for a more informed public debate on these issues and bringing data to the table for better informed policy making.

Today the main way users can protect themselves against online tracking is by installing browser privacy tools such as Ghostery, AdBlock Plus, or uBlock Origin. Measurement research helps improve these tools — sometimes by finding entirely new categories of tracking and sometimes by finding new trackers employing known types of tracking. The block lists used by these tools were compiled by a laborious manual process, but automated methods allow us to find new trackers quickly and efficiently. The data released by our project finds its way into these tools [43].

Even so, today's tracker-blocking tools have important limitations: they block many benign URLs and break the functionality of a significant fraction of sites. In ongoing research, we are looking at the possibility that these tools can be built in a radically different manner: using machine learning to automatically learn the difference between tracking and non-tracking *behavior* instead of *actors*. [3] The machine-learning classifier would be trained on our web-scale census datasets, and the browser tool would download this classifier instead of lists of trackers. There's a bit of poetic justice in using the tools of Big Data and machine learning, which are used by the tracking industry, to instead protect users against tracking.

5 Implications for Regulating Privacy

The web browser as a privacy regulator. As we saw in Section 2.3, the web browser mediates the user's interaction with web pages and trackers, and so browser vendors have considerable power over the state of online tracking. Vendors have been aware of this, but most have traditionally tended to remain "neutral". For example, browsers other than Safari have avoided blocking third-party cookies by default for this reason.

We view this stance as misguided. The web is so complex that the browser's defaults, user interface, and extension interface have an inevitable and tremendous impact on users' privacy outcomes. In the language of *Nudge*, the browser is a "choice architecture" and hence cannot be neutral [45]. In practice, attempts at neutrality have the effect of simply leaving in place the results of historical accidents. For example, the web standard explicitly leaves deciding how to handle third-party cookies to the browser, and most browsers made their original permissive decisions in a historical context where the privacy implications were not as clear as they are today.

[3] The Privacy Badger tool (https://www.eff.org/privacybadger) works somewhat like this, but it uses hard-coded heuristics instead of machine learning.[44]

More recently, this attitude has been changing. Apple yielded to user demands to enable content blocking on Safari for iOS [46], Chrome is deliberating removing the filesystem API due to privacy-infringing use [47], and Microsoft enabled the Do Not Track signal by default in Internet Explorer 10 [48].[4] Firefox has been leading the charge, removing the Battery API due to abuse [49], enabling tracking protection in private browsing mode [50], and experimenting with other advanced privacy features [51].

We urge browser vendors to embrace their role as regulators of web privacy. There are important ongoing battles that pit consumer privacy against commercial interests, and browsers cannot and should not avoid taking a side. This will be especially tricky for browsers made by companies involved in online advertising. But browsers have the opportunity, through innovative technology, to steer the debate away from its current framing as a zero-sum game.

Open standards and privacy. The above discussion pertains to the baseline level of privacy on the web, i.e., the privacy outcome for the hypothetical average user. However, more technically skilled users may benefit from enabling optional browser features (including those squirreled away in "power user" interfaces) and installing and configuring browser extensions. In general, in an open platform, technologically savvy users are in a dramatically better position to protect their privacy and security, whereas in a closed platform, privacy and security outcomes are relatively uniform. Neither model is strictly better for privacy, but they do result in very different outcomes.

Inequality of privacy outcomes based on technical skill is worrying by itself, but also because such skill correlates with historically advantaged groups. It is a concern for researchers: new privacy technologies favor the tech savvy, so privacy research may *exacerbate* inequality in privacy outcomes unless combined with outreach efforts. This bias towards the tech savvy may also lead to a focus on technologies that are simply unsuitable for mainstream use, such as PGP. Inequality is especially a concern for the open-source model of privacy tool development, because in most cases there is no funding for usability testing or user education. There is a clear role for journalists, privacy activists, and civil society organizations to bridge the gap between developers and users of privacy tools. Consumer protection agencies could also play a role.

Policy makers should take note of the differences between open vs. closed platforms. As new platforms for data collection (such as the Internet of Things) take shape, it will be important to understand whether they lean open or closed. Open platforms present challenges for regulation since there isn't a natural point of leverage, jurisdictional boundaries are harder to enforce, and the "long tail" of innovation makes enforcement difficult to scale. Given these challenges, one avenue for regulators is to complement technical measures, such as by clamping down on circumvention of cookie blocking [52]. We discuss two more approaches in the following two subsections.

[4] The Do Not Track standard itself lacks any teeth because of the failure of attempts at legislation or regulation to give it enforceable meaning.

The market for lemons and first-party accountability. Many web publishers, as we have noted, have little awareness of the tracking on their own sites and the implications of it. The lack of oversight of third parties by publishers is a problem for privacy. This is most clearly seen in terms of the economic view that we used earlier. In a well-functioning market, many consumers will consider privacy (as one of several factors) in picking a product, service, website, etc. But in the case of online tracking, privacy failings are seen as the responsibility of third parties rather than publishers. Users don't interact directly with third parties, and therefore have no way to exercise their preferences in the marketplace.

We think this needs to change. When tracking is reported in the press, journalists should seek to make first parties accountable, and not just third parties. Similarly, privacy laws should make first parties primarily responsible for privacy violations. These will shift incentives so that first parties will start to have oversight of the tracking that happens on their domains. One of the very few examples that already embodies this principle is the US Children's Online Privacy Protection Act (COPPA), especially the Federal Trade Commission's COPPA rule [53] and recent enforcement actions against child-directed websites [54].

This transition won't be easy. Due to the financial struggles of publishers, regulators are loath to impose additional compliance burdens on them. But measurement tools can help. As part of WebTAP, we are building a publisher dashboard for a website operator to understand the tracking technologies in use on their own domain. Combined with the right technologies, the shift in incentives can in fact be a boon for publishers, who currently lack adequate information not just about tracking technologies but also about how tracking translates into revenue. In other words, our suggestion is to help shift the balance of power (and, with it, responsibility) from third parties to publishers.

Design for measurement. Finally, our work suggests that policymakers have a lightweight way to intervene to improve privacy: by requiring service providers to support the ability of external researchers to measure and audit privacy. This could be as straightforward as the creation of APIs (application programming interfaces) for external measurement, to help automate studies that would otherwise require arduous manual effort. At the very least, policymakers should work to remove existing legal barriers to measurement research. Recently, the American Civil Liberties Union, together with academics, researchers, and journalists, have challenged the constitutionality of the Computer Fraud and Abuse Act on the grounds that it prevents uncovering racial discrimination online [55]. We welcome this development.

Transparency through external oversight is a valuable complement to — and sometimes more effective than — transparency through notice (i.e., privacy policies), for several reasons. Most importantly, external oversight doesn't require trusting the word of the company. Often, leaks of personally-identifiable information (PII) or discriminatory effects of algorithms are introduced unintentionally into code, and measurement offers a way to discover these. Finally, measurement-based transparency can provide a precise and quantifiable view of privacy.

6 The Future of Transparency Research

Online tracking has proved an amenable target for large-scale measurement. Replicating the success of the WebTAP project in other domains will be much harder. For example, while there have been several interesting and important studies of privacy on smartphones and mobile devices, they don't reach the same scale and completeness, since app platforms are not as programmable as browsers [56, 57, 58] .

So far, WebTAP has looked primarily at data collection and data flows, and not nearly as much at uncovering bias and discrimination — and more generally, discovering how personal data is being used behind the scenes by algorithms. Here again it becomes harder to scale, because getting insights into personalization on a single website might require hundreds or thousands of observations. It also requires developing statistical techniques for establishing correlation or causation. Research teams at Columbia University and Carnegie Mellon University have recently made progress on this problem [59, 60, 61, 62].

The most challenging scenarios for privacy measurement are also among the most important: those that involve the physical world. Think of the "Internet of Things" monitoring your activities in your home; tracking by analytics firms in shopping malls, based on WiFi and other emanations from your smartphone; apps that track your location and other activities; cross-device tracking and onboarding discussed in Section 4.3. The difficulty for researchers, of course, is that we can't quite automate the real world, at least not yet.

Researchers are adapting to these challenges in various ways. "Crowdsourcing" of data from different users' devices has resulted in important findings on privacy, price discrimination, and so on [8]. Sometimes it is possible to manipulate a fake user's location automatically [63, 64]. Engineers have created tools called "monkeys" that can simulate a user clicking through and exploring a smartphone app [65]. These were developed for finding bugs, but can also be used to automatically detect if the app phones home with personal data [66]. Computer science techniques such as static analysis and dynamic analysis allow analyzing or running an app in a simulated setting to understand its behavior [67]. Monitoring network communications generated by smartphones has proved powerful, especially combined with techniques to peek into encrypted traffic [68, 69]. Occasionally, researchers have rolled up their sleeves and conducted experiments manually in the absence of automated tools [70]. Finally, companies have sometimes been forthcoming in making provisions for researchers to study their systems.

That last point is important. Privacy research has too often been adversarial, and efforts by companies to be transparent and work with external researchers should be encouraged and rewarded. In conjunction, we need a science of designing systems to be transparent from the ground up. In the last few years, the Fairness, Accountability, and Transparency in Machine Learning ("FAT-ML") research community has made progress in developing algorithms that respect norms of transparency and non-discrimination.

While this progress is laudable, it appears that we have a long way to go in figuring out how to develop technology, especially machine learning, that respects

our societal norms. A recent Princeton research paper (co-authored by Narayanan) looked for bias in machine learning — specifically, in a state-of-the-art technique called "word embeddings" that provide an algebraic representation of words that are easy for computers to manipulate [71]. The authors started from the "Implicit Association Test," a standard method in psychology to test human biases, and developed a version of it for word embeddings. They were able to replicate in the machine learning model every bias they tested that's been documented in humans, including racial and gender biases.

In other words, the underlying model of language used by the machine for a wide variety of tasks (such as language translation) is intrinsically biased. The authors argue that since these models are trained on text from the web written by humans, the machine inevitably absorbs the entire spectrum of human biases and prejudices, along with the rest of language, meaning and semantics. It is impossible to learn one without learning the other. This means that if we want to avoid enshrining our historical prejudices in our algorithms, we have to fundamentally re-examine the reigning paradigm of training machines on human knowledge and intelligence, and that will require a long-term research program.

Conclusion. The technology industry innovates at breakneck pace. But the more data-driven algorithms encroach into our lives, the more their complexity and inscrutability becomes problematic. A new area of empirical research seeks to make these systems more transparent, study their impact on society, and enable a modicum of external oversight. The Princeton Web Transparency and Accountability Project has focused on a small but important piece of this puzzle: third-party online tracking. By exploiting the open nature of web technologies, we have been able to track the trackers in an automated, large-scale, continual fashion, and to conduct a comprehensive study of tracking technologies used online.

An exciting but challenging future lies ahead for transparency research. Studying domains and systems that are less amenable to automated measurement will require various creative ideas. We hope these research findings will shape public policy, just as environmental policy is shaped by research examining the impact of human activities. Perhaps the greatest need is to develop a science of building data-driven algorithms in an ethical and transparent fashion from the ground up. Perhaps in the future algorithmic systems will even be built to explicitly support external measurement and oversight.

Acknowledgement. Numerous graduate and undergraduate students and collaborators have contributed to the WebTAP project and to the findings reported here. In particular, Steven Englehardt is the primary student investigator and the lead developer of the OpenWPM measurement tool. We are grateful to Brian Kernighan, Vincent Toubiana, and the anonymous reviewer for useful feedback on a draft.

WebTAP is supported by NSF grant CNS 1526353, a grant from the Data Transparency Lab, and by Amazon AWS Cloud Credits for Research.

References

[1] Crevier D (1993) AI: The tumultuous history of the search for artificial intelligence. Basic Books, Inc.

[2] Engle Jr RL, Flehinger BJ (1987) Why expert systems for medical diagnosis are not being generally used: a valedictory opinion. Bulletin of the New York Academy of Medicine 63(2):193

[3] Vance A (2011) This tech bubble is different. http://www.bloomberg.com/

[4] Angwin J (2016) Machine bias: Risk assessments in criminal sentencing. ProPublica https://www.propublica.org/

[5] Levin S (2016) A beauty contest was judged by AI and the robots didn't like dark skin. https://www.theguardian.com/

[6] Solove DJ (2001) Privacy and power: Computer databases and metaphors for information privacy. Stanford Law Review pp 1393–1462

[7] Marthews A, Tucker C (2015) Government surveillance and internet search behavior. Available at SSRN 2412564

[8] Hannak A, Soeller G, Lazer D, Mislove A, Wilson C (2014) Measuring price discrimination and steering on e-commerce web sites. In: Proceedings of the 2014 conference on internet measurement conference, ACM, pp 305–318

[9] Calo R (2013) Digital market manipulation. University of Washington School of Law Research Paper 2013-27 DOI 10.2139/ssrn.2309703

[10] Mayer JR, Mitchell JC (2012) Third-party web tracking: Policy and technology. In: 2012 IEEE Symposium on Security and Privacy, IEEE, pp 413–427

[11] Angwin J (2010) The web's new gold mine: Your secrets. ProPublica http://www.wsj.com/

[12] Lerner A, Simpson AK, Kohno T, Roesner F (2016) Internet jones and the raiders of the lost trackers: An archaeological study of web tracking from 1996 to 2016. In: 25th USENIX Security Symposium (USENIX Security 16)

[13] Laperdrix P, Rudametkin W, Baudry B (2016) Beauty and the beast: Diverting modern web browsers to build unique browser fingerprints. In: 37th IEEE Symposium on Security and Privacy (S&P 2016)

[14] Eckersley P (2010) How unique is your web browser? In: International Symposium on Privacy Enhancing Technologies Symposium, Springer, pp 1–18

[15] Acar G, Van Alsenoy B, Piessens F, Diaz C, Preneel B (2015) Facebook tracking through social plug-ins. Technical report prepared for the Belgian Privacy Commission https://securehomes.esat.kuleuven.be/ gacar/fb_tracking/fb_plugins.pdf

[16] Starov O, Gill P, Nikiforakis N (2016) Are you sure you want to contact us? quantifying the leakage of pii via website contact forms. Proceedings on Privacy Enhancing Technologies 2016(1):20–33

[17] Krishnamurthy B, Naryshkin K, Wills C (2011) Privacy leakage vs. protection measures: the growing disconnect. In: Proceedings of the Web, vol 2, pp 1–10

[18] Su J, Shukla A, Goel S, Narayanan A (2017) De-anonymizing web browsing data with social networks, manuscript

[19] Barocas S, Nissenbaum H (2014) Big data's end run around procedural privacy protections. Communications of the ACM 57-11:31-33
[20] Shilton K, Greene D (2016) Because privacy: defining and legitimating privacy in ios development. IConference 2016 Proceedings
[21] Storey G, Reisman D, Mayer J, Narayanan A (2016) The future of ad blocking: Analytical framework and new techniques, manuscript
[22] Narayanan A (2016) Can Facebook really make ads unblockable? https://freedom-to-tinker.com/
[23] Storey G (2016) Facebook ad highlighter. https://chrome.google.com/webstore/detail/facebook-ad-highlighter/mcdgjlkefibpdnepeljmlfkbbbpkoamf?hl=en
[24] Reisman D (2016) A peek at A/B testing in the wild. https://freedom-to-tinker.com/
[25] Acar G, Juarez M, Nikiforakis N, Diaz C, Gürses S, Piessens F, Preneel B (2013) Fpdetective: dusting the web for fingerprinters. In: Proceedings of the 2013 ACM SIGSAC conference on Computer & communications security, ACM, pp 1129–1140
[26] Englehardt S, Narayanan A (2016) Online tracking: A 1-million-site measurement and analysis. In: Proceedings of the 2016 ACM SIGSAC conference on Computer & communications security
[27] Selenium HQ (2016) Selenium browser automation faq. https://code.google.com/p/selenium/wiki/FrequentlyAskedQuestions
[28] Acar G, Eubank C, Englehardt S, Juarez M, Narayanan A, Diaz C (2014) The web never forgets. Proceedings of the 2014 ACM SIGSAC Conference on Computer and Communications Security - CCS '14 DOI 10.1145/2660267.2660347
[29] Mowery K, Shacham H (2012) Pixel perfect: Fingerprinting canvas in html5. Proceedings of W2SP
[30] (Valve) VV (2016) Fingerprintjs2 — modern & flexible browser fingerprinting library, a successor to the original fingerprintjs. https://github.com/Valve/fingerprintjs2
[31] Olejnik Ł, Acar G, Castelluccia C, Diaz C (2015) The leaking battery. In: International Workshop on Data Privacy Management, Springer, pp 254–263
[32] Englehardt S, Narayanan A (2016) Online tracking: A 1-million-site measurement and analysis. Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security - CCS '16
[33] Doty N (2016) Mitigating browser fingerprinting in web specifications. https://w3c.github.io/fingerprinting-guidance/
[34] Soltani A, Peterson A, Gellman B (2013) NSA uses Google cookies to pinpoint targets for hacking. https://www.washingtonpost.com/
[35] Englehardt S, Reisman D, Eubank C, Zimmerman P, Mayer J, Narayanan A, Felten EW (2015) Cookies that give you away. Proceedings of the 24th International Conference on World Wide Web - WWW '15 DOI 10.1145/2736277.2741679

[36] Angwin J (2016) Google has quietly dropped ban on personally identifiable web tracking. ProPublica https://www.propublica.org
[37] Reitman R (2012) What actually changed in Googles privacy policy. Electronic Frontier Foundation https://www.eff.org
[38] Simonite T (2015) Facebooks like buttons will soon track your web browsing to target ads. MIT Technology Review https://www.technologyreview.com/
[39] Federal Trade Commission (2015) Cross-device tracking. https://www.ftc.gov/news-events/events-calendar/2015/11/cross-device-tracking
[40] Maggi F, Mavroudis V (2016) Talking behind your back attacks & countermeasures of ultrasonic cross-device tracking, https://www.blackhat.com/docs/eu-16/materials/eu-16-Mavroudis-Talking-Behind-Your-Back-Attacks-And-Countermeasures-Of-Ultrasonic-Cross-Device-Tracking.pdf, blackhat
[41] Angwin J (2014) Why online tracking is getting creepier. ProPublica https://www.propublica.org/
[42] Vallina-Rodriguez N, Sundaresan S, Kreibich C, Paxson V (2015) Header enrichment or isp enrichment? Proceedings of the 2015 ACM SIGCOMM Workshop on Hot Topics in Middleboxes and Network Function Virtualization - HotMiddlebox '15 DOI 10.1145/2785989.2786002
[43] Disconnect (2016) Disconnect blocks new tracking device that makes your computer draw a unique image. https://blog.disconnect.me/disconnect-blocks-new-tracking-device-that-makes-your-computer-draw-a-unique-image/
[44] Foundation EF (2016) Privacy badger. https://www.eff.org/privacybadger
[45] Thaler RH, Sunstein CR (2008) Nudge: improving decisions about health, wealth, and happiness. Yale University Press
[46] Fleishman G (2015) Hands-on with content blocking safari extensions in ios 9. Macworld http://www.macworld.com/
[47] Blink, Chromium (2016) Owp storage team sync. https://groups.google.com/a/chromium.org/forum/#!topic/blink-dev/CT_eDVIdJv0
[48] Lynch B (2012) Do not track in the windows 8 setup experience - microsoft on the issues. Microsoft on the Issues https://blogs.microsoft.com/
[49] Hern A (2016) Firefox disables loophole that allows sites to track users via battery status. The Guardian https://www.theguardian.com/
[50] Mozilla (2015) Tracking protection in private browsing. https://support.mozilla.org/en-US/kb/tracking-protection-pbm
[51] Mozilla (2016) Security/contextual identity project/containers. https://wiki.mozilla.org/Security/Contextual_Identity_Project/Containers
[52] Federal Trade Commission (2012) Google will pay $22.5 million to settle FTC charges it misrepresented privacy assurances to users of apple's safari internet browser. https://www.ftc.gov/news-events/press-releases/2012/08/google-will-pay-225-million-settle-ftc-charges-it-misrepresented

[53] Federal Trade Commission (2016) Children's online privacy protection rule ("coppa"). https://www.ftc.gov/enforcement/rules/rulemaking-regulatory-reform-proceedings/childrens-online-privacy-protection-rule
[54] New York State Office of the Attorney General (2016) A.G. schneiderman announces results of "operation child tracker," ending illegal online tracking of children at some of nation's most popular kids' websites. http://www.ag.ny.gov/press-release/ag-schneiderman-announces-results-operation-child-tracker-ending-illegal-online
[55] American Civil Liberties Union (2016) Sandvig v. Lynch. https://www.aclu.org/legal-document/sandvig-v-lynch-complaint-0
[56] Eubank C, Melara M, Perez-Botero D, Narayanan A (2013) Shining the floodlights on mobile web tracking a privacy survey. http://www.w2spconf.com/2013/papers/s2p2.pdf
[57] CMU CHIMPS Lab (2015) Privacy grade: Grading the privacy of smartphone apps. http://www.privacygrade.org
[58] Vanrykel E, Acar G, Herrmann M, Diaz C (2016) Leaky birds: Exploiting mobile application traffic for surveillance. Financial Cryptography and Data Security 2016
[59] Lécuyer M, Ducoffe G, Lan F, Papancea A, Petsios T, Spahn R, Chaintreau A, Geambasu R (2014) Xray: Enhancing the webs transparency with differential correlation. In: 23rd USENIX Security Symposium (USENIX Security 14), pp 49–64
[60] Lecuyer M, Spahn R, Spiliopolous Y, Chaintreau A, Geambasu R, Hsu D (2015) Sunlight: Fine-grained targeting detection at scale with statistical confidence. In: Proceedings of the 22nd ACM SIGSAC Conference on Computer and Communications Security, ACM, pp 554–566
[61] Tschantz MC, Datta A, Datta A, Wing JM (2015) A methodology for information flow experiments. In: 2015 IEEE 28th Computer Security Foundations Symposium, IEEE, pp 554–568
[62] Datta A, Sen S, Zick Y (2016) Algorithmic transparency via quantitative input influence. In: Proceedings of 37th IEEE Symposium on Security and Privacy
[63] Chen L, Mislove A, Wilson C (2015) Peeking beneath the hood of uber. In: Proceedings of the 2015 ACM Conference on Internet Measurement Conference, ACM, pp 495–508
[64] Valentino-Devries J, Singer-Vine J, Soltani A (2012) Websites vary prices, deals based on users information. Wall Street Journal 10:60–68
[65] Guide ASU (2016) Ui/application exerciser monkey. https://developer.android.com/studio/test/monkey.html
[66] Rastogi V, Chen Y, Enck W (2013) Appsplayground: automatic security analysis of smartphone applications. In: Proceedings of the third ACM conference on Data and application security and privacy, ACM, pp 209–220
[67] Enck W, Gilbert P, Han S, Tendulkar V, Chun BG, Cox LP, Jung J, McDaniel P, Sheth AN (2014) Taintdroid: an information-flow tracking system for realtime privacy monitoring on smartphones. ACM Transactions on Computer Systems (TOCS) 32(2):5

[68] Ren J, Rao A, Lindorfer M, Legout A, Choffnes D (2015) Recon: Revealing and controlling privacy leaks in mobile network traffic. arXiv preprint arXiv:150700255

[69] Razaghpanah A, Vallina-Rodriguez N, Sundaresan S, Kreibich C, Gill P, Allman M, Paxson V (2015) Haystack: in situ mobile traffic analysis in user space. arXiv preprint arXiv:151001419

[70] Sweeney L (2013) Discrimination in online ad delivery. Queue 11(3):10

[71] Caliskan-Islam A, Bryson J, Narayanan A (2016) Semantics derived automatically from language corpora necessarily contain human biases. Arxiv https://arxiv.org/abs/1608.07187