Third-Party Net Stalk: Policy and Technology

S.Vanitha¹, U.Palani² A.Arthibala³

Assistant Professor, Associate Professor
IFET College of Engineering, Villupuram

I. INTRODUCTION

The net has evolved to facilitate development and delivery of webpages composed of content from multiple websites. HTML, JavaScript, and CSS impose no restrictions on a webpage including elements from, or even delegating complete control to, a wholly unrelated website.¹ These design choices have contributed to a host of well-known and well-studied security vulnerabilities, including cross-site scripting (XSS) [2] and cross-site request forgery (CSRF or XSRF) [3], [4], that enable an unauthorized and unrelated “third-party” website to retrieve information from or perform actions on the “first-party” website that the user has voluntarily interacted with. This paper examines the privacy implications of the opposite case—where a first-party website authorizes a third-party website to learn about its users (Figure 1).² Third-party services bring tremendous value to the web: they enable first-party websites to trivially implement advertising, analytics, social network integration, and more. But they also give rise to privacy concerns: over the past several years, researchers, civil society organizations, and policymakers have called attention to the increasing trend of third-party websites recording and analysing users’ browsing activities across unrelated first-party websites (“third-party web tracking” or “tracking” for short).³ This paper is intended to comprehensively familiarize computer security and privacy researchers with current policy and technology research on third-party web tracking. Much of the discussion is based on recent results from a new dynamic web measurement platform, Fourth Party. We begin by presenting by Fourth Party. The remainder of the paper is organized into two parts on third-party web tracking: one on policy, and one on technology. The policy part opens by reviewing why third-party web tracking gives rise to privacy concerns and ways in which policy might be structured to address those concerns. It then provides an overview of regulation and self-regulation in the U.S. and EU, explaining the current governmental and business approaches to mitigating privacy issues in third-party web tracking. The discussion next turns to tracking business models and trends. A final section frames the economic debate on third-party web tracking and notes gaps in the current literature.

The technology part starts by surveying stateful and stateless technologies that can be used to correlate users’ activities across websites. It next provides an overview of technologies that enable the delivery of third-party services with lessened privacy risk. Last, it reviews the user choice and self-help technologies presently available, including opt-out cookies, blocking, and Do Not Track. This paper has a secondary aim. Debates on how to respond to third-party web tracking are now occurring every day in Washington and Brussels. We hope that by systematizing knowledge on third-party web tracking for the computer security and privacy community, we will ensure that it is best able to assist policymakers in developing solutions that adequately balance privacy, commerce, and a thriving web.

II. FOURTHPARTY

• Why Web Measurement?

We have found several advantages to placing web measurement at the centre of our methodology. Web measurement provides objective, reliable evidence that both furthers public understanding and establishes a sound basis for policymaking. Second, web measurement is fast. Many claims about specific tracking practices can be supported or rebutted with mere hours of web measurement work.

Web measurement facilitates longitudinal study. Often the very same hardware and software can be reused to collect and analyze data even years apart. Last, web measurement can often be automated. Once a generic measurement tool has been built, it can be trivially applied to millions of websites.

• Design Principles

Prior work on third-party web tracking has largely taken one of three approaches to measurement: monitor network traffic (e.g. [10], [11], [12], [13]), manually inspect browser state (e.g. [8], [14]), or develop a custom tool for a specific measurement task (e.g. [15], [16], [17]). We developed FourthParty around three design principles that improve on these approaches.

1) General-purpose instrumentation: By implementing comprehensive instrumentation and logging only once, Fourth Party avoids the need for many purpose-built tools, decreases duplication of effort, and trims development time.

2) Production web browser: Building on a production browser allows reuse of existing add-ons, including for automation, and closely emulates real-world browsing.

3) Standardized log format: A standardized, easy-to-manipulate log format facilitates data sharing and cuts back on redundant data gathering.

• Implementation

We implemented FourthParty as an extension to Mozilla Firefox. It currently instruments the browser APIs for HTTP traffic, DOM windows, cookies, and resource loads. Fourth-Party also instruments JavaScript API calls on the window, navigator, and screen objects using getters, setters, and ECMAScript proxies [18]. All events are logged to a SQLite database. On many pages FourthParty does not perceptibly slow down Firefox; on highly dynamic pages, it can increase page load time by roughly 2-3x.

We plan to make substantial performance improvements in an upcoming revision.

• Analysis with FourthParty

Analyzing FourthParty data is fast. All of the FourthParty results presented in this paper were generated with Python scripts, most of which took seconds to execute on a con-sumer notebook with databases

Figure 1. Third-party advertising, social, and video content on the New York Times website. Analytics content is not visible.
including visits to thousands of popular websites. Analyzing FourthParty data is also easy for researchers who are already familiar with SQL syntax. For example, a query that counts Google Analytics reports with an anonymized instruction (see Section VIII-B) is just:

```sql
select count(*) from http_requests where url like '%sgoogle-analytics.com/utm.gif?p=%'
```

See http://fourthparty.info for the FourthParty source and related resources.

We used FourthParty to conduct many of our studies. We present in this paper, including on information leakage (Section III-B3), tracking technologies (Section VII), and blocking tool effectiveness (Section IX-B).

## Third-Party Web Tracking Policy

### III. PRIVACY PROBLEMS

This section reviews the privacy implications of third-party web tracking and notes the range of policy responses. The discussion proceeds in four phases. First, it details the browsing history information that is available to third parties and how that information is identifiable. Second, it explains how third-party web tracking could harm users. Third, it reviews survey results consistently showing that users would prefer to not be tracked. Last, it details the policy positions that various stakeholders have adopted in response.

#### Information Available

Web browsing history is intrinsically linked to personal information. The pages a user visits can reveal her location, interests, purchases, employment status, sexual orientation, financial challenges, medical conditions, and more. Examining individual page loads is often adequate to draw many conclusions about a user; analyzing patterns of activity allows yet more inferences.

When a first-party page embeds third-party content, the third-party website is ordinarily made aware of the URL of the first-party page through an HTTP referrer or equivalent. If the page embeds a script tag from a third party, the third party will also often learn the webpage's title from document.title. Some first parties will voluntarily transmit even more information.

Collection of sensitive personal information is not a hypothetical concern. In mid-2011 we discovered that an advertising network, Epic MarketPlace, had publicly exposed its interest segment data, offering a rare glimpse of what third-party trackers seek to learn about users [7]. User segments included menopause, getting pregnant, repairing bad credit, and debt relief. Several months later we found that the free online dating website OkCupid was sending to the data provider Lotame how often a user drinks, smokes, and does drugs [19]. When Krishnamurthy et al. [10] tested search queries on ten popular health websites, they found a third party learned of the user's query on nine of them.

#### Identifiability

A web browsing history is often personally identifiable or identifiable. Narayanan [20] recently proposed a taxonomy of five ways in which a pseudonymous browsing history might become identifiable. Note that pseudonymity is quite fragile in protecting identity: discovering a user’s identity once in a pseudonymous system is sufficient to also identify past and future interactions with the user.

1) The third party is also a first party: The third party may be a first party in another context, where the user voluntarily provided her identity. Facebook, for example, has over 800 million users and enforces a requirement that users provide their real name to the service. When a page includes a third-party Facebook social widget, Facebook identifies the user to personalize the widget.

2) A first party sells the user’s identity: Some first-party websites voluntarily provide (“leak”) a user’s identity to third parties for pay. Some have even made a business model of it, usually appearing as a free sweepstakes or quiz. Several advertising data providers (e.g. Datalogix [21]) buy identifying information, retrieve the user’s dossier from an offline consumer database, and use it to target advertising.

3) A first party unintentionally provides identity: If a website puts identifying information in a URL or page title, it may unintentionally leak the information to third parties. In a 2011 paper [10], Krishnamurthy et al. examined signup and interaction with 120 popular sites for information leakage to third parties. They reported that an aggregate of 48% leaked a user identifier in a Request-URI or referrer.

Using a similar methodology, we examined identifying information leakage on the Quantcast U.S. top 250 websites [19]. We were able to test signup and interaction with 185 of the sites; we found that a username or user ID was sent to a domain with a different public suffix +1 (PS+1) on 113 (61%) of the websites in our sample. The five most frequent recipients and most prolific senders of username and user ID are presented in Table I and Table II respectively.

In the majority of instances the username or user ID was part of a user profile URL or page title. A better practice

### TABLE I

<table>
<thead>
<tr>
<th>Third-Party PS+1</th>
<th>Websites Leaking Username or ID</th>
</tr>
</thead>
<tbody>
<tr>
<td>scorecardresearch.com</td>
<td>81 (44%)</td>
</tr>
<tr>
<td>google-analytics.com</td>
<td>78 (42%)</td>
</tr>
<tr>
<td>quantserve.com</td>
<td>63 (34%)</td>
</tr>
<tr>
<td>doubleclick.net</td>
<td>62 (34%)</td>
</tr>
<tr>
<td>facebook.com</td>
<td>45 (24%)</td>
</tr>
</tbody>
</table>

#### Table II

<table>
<thead>
<tr>
<th>Popular Websites Leaking Username and ID.</th>
</tr>
</thead>
<tbody>
<tr>
<td>First-Party PS+1</td>
</tr>
<tr>
<td>-----------------</td>
</tr>
<tr>
<td>rotten tomatoes.com</td>
</tr>
<tr>
<td>cafemom.com</td>
</tr>
<tr>
<td>lyricsmode.com</td>
</tr>
<tr>
<td>ivillage.com</td>
</tr>
<tr>
<td>livejournal.com</td>
</tr>
</tbody>
</table>

would be to use a single URL for all users viewing their own profile, e.g. http://example.com/self, and to never include the username or user ID in the page title. Several of the sites we contacted were willing to implement these fixes, but many more preferred the functionality, con-venience, and aesthetic of a username or user ID in URLs. It seems quite likely the practice will persist indefinitely among even the most popular sites.

We also observed other forms of identifying information leak. For example:

1) Viewing a local ad on the Home Depot website sent the user’s first name and email address to 13 companies.

2) Entering the wrong password on the Wall Street Journal website sent the user’s email address to 7 companies.

3) Changing user settings on the video-sharing site Meta-cafe sent first name, last name, birthday, email address, physical address, and phone numbers to 2 companies.

In all of these cases the identifying information was included as a parameter in a first-party URL. The better practice is to send identifying information as part of a POST request body so it will not inadvertently leak to third parties.

4) The third party uses a security exploit: A third party may exploit a cross-site security vulnerability on a first-party website to learn the user’s identity. Narayanan has shown how inadequate frame busting can facilitate identifying a user [27]. Huang and Jackson more recently demonstrated practical user identification through Facebook and Twitter sharing widget clickjacking [28].

5) Re-identification: The third party could match pseudonymous browsing histories against identified datasets to re-identify them, much like Narayanan and Shmatikov did with the Netflix Prize dataset [29] and the Flickr and Twitter social graphs [30], and Acquisti et al. did more recently with personal photos on a dating site [31]. A third party might, for example, compare browsing activity to the times and locations of links publicly shared by Twitter users.

#### Possible Harms

The risk of harm to consumers from web tracking arises from myriad potential scenarios. Each particular scenario may have a low probability of occurring. But the chance of some scenarios occurring is substantial, especially when considered over time and across many companies.

When considering harmful web tracking scenarios, we find it helpful to focus on four variables. First, an actor that causes harm to a consumer. The actor might, for example, be an authorized employee, malicious employee, competitor, acquirer, hacker, or government agency. Second, a means of access that enables the actor to use tracking data. The data might be voluntarily transferred, sold, stolen, misplaced, or accidentally distributed. Third, an action that harms the consumer. The action could be, for example, publication, a less favorable offer, denial of a benefit, or termination of employment. Last, a particular harm that is inflicted. The harm might be physical, psychological, or economic.

The countless combinations of these variables result in countless
possible bad outcomes for consumers. To ex-emptify our thinking, here is one commonly considered scenario: A hacker (actor) breaks into a tracking company (means of access) and publishes its tracking information (action), causing some embarrassing fact about the consumer to become known and inflicting emotional distress (harm). Risks associated with third-party tracking are heightened by the lack of market pressure to exercise good security and privacy practices. If a first-party website is untrustworthy, users may decline to visit it. But, since users are unaware of the very existence of many third-party websites, they cannot reward responsible sites and penalize irresponsible sites.10

- User Preferences

User surveys have consistently shown opposition to third parties collecting and using browsing activity. A 2009 representative U.S. phone survey by Turow et al. [33] found that 87% of respondents would not want advertising based on tracking. In an unrepresentative 2010 survey of Amazon Mechanical Turk users by McDonald and Cranor [34], only 45% of respondents wanted to be shown any ads that had been tailored to their interests. A December 2010 USA Today/Gallup poll [35] reported 67% of respondents thought behavioural targeting should be outright illegal. In a mid-2011 representative U.S. online survey by Truste and Harris Interactive [36], 85% of respondents said they would not consent to tracking for ad targeting, and 78% said they would not consent to tracking for website analytics. Finally, a 2012 representative telephone survey by Pew Research found that 98% of respondents were “not okay” with behavioural advertising [37].

One area for future survey work is in disaggregating user preferences about collection of tracking data from preferences about specific uses of tracking data. The survey literature has largely (but not entirely) focused on behavioral advertising, which can conflate data collection and use. Another area for future research is preference balancing. All of the above studies examined user preferences inde-pendent of economic considerations; there remains a need for work that more directly examines the economic tradeoffs users would make to be or not be tracked.

- Policy Views

Policy views on third-party web tracking vary substan-tially. All stakeholders agree that consumers should have some degree of control over web tracking, but there are many points of disagreement on specifics. What should consumers be able to control? Many poli-cymakers and advocates believe consumers should have control over the collection of web tracking information. Online advertising trade groups have argued that control should only extend to specific uses of data. What should the default be? EU policymakers believe no tracking should be the default [38]; advertising trade groups argue that tracking should be the default [39].

Who should design the choice mechanism? Advertis ing trade groups would like to control choice mechanism design [39]. Many policymakers and advocates believe the browser vendors should retain design responsibility. Views on web tracking policy are, of course, shaped by underlying priorities. Some, particularly consumer advocates and EU policymakers, view online privacy as a fundamental human right. Others, including many researchers and U.S. policymakers, see consumer choice about tracking privacy risks as a means to maximize welfare.11 Mozilla [40], [41] takes the position that giving consumers a choice about tracking is itself a policy goal. Third-party websites and advertising trade groups largely defend current practices with arguments rooted in welfare—that the subsidy to content outweighs consumer privacy risks—and economic rights.

IV. REGULATION AND SELF-REGULATION

Third-party, user-tracking has, unusually, largely ex-isted in a regulatory vacuum. The following subsections detail limits imposed by U.S. and EU law, as well as the online advertising industry’s self-regulatory programs.

- United States

The Federal Trade Commission (FTC) is the leading federal regulatory and law enforcement agency for consumer protection. The FTC has narrowly circumscribed general statutory authority: it can only prevent business practices that are either “unfair” or “deceptive” under 15 U.S.C. § 45. On tracking issues the agency has generally relied on its deception authority, where a company breaches an express representation it has made to consumers.12 The FTC almost always settles a company’s first violation with a consent order and slight (if any) payment; though not directly financially punitive, business are loath to endure the expense, burden, and negative publicity of a federal law enforcement action. A subsequent violation of a consent order can result in significant monetary penalties.

- Online Advertising Self-Regulation

The online advertising industry has largely harmonized self-regulatory efforts in the U.S. (the Network Advertising Initiative, NAI [60] and the Digital Advertising Alliance, DAA [61]) and the EU (the Interactive Advertising Bureau Europe, IAB Europe [62]). All three programs impose the same consumer choice requirement: participating companies must allow users to opt out of behavioural advertising, that is, ad targeting based on tracking. Note that this is a choice about one particular use of data; collection and other uses of third-party tracking data are unaffected.13 Participation in self-regulation has fluctuated with waxing and waning government scrutiny [65]. At present most of the largest online advertising and analytics companies participate, and most of the smaller ones do not.

- European Union

The 2002 ePrivacy Directive, 2002/58/EC, mandated that websites enable users to opt out of having information stored in their browser, except as “strictly necessary” to provide service “explicitly requested” by the user. In practice the directive has had little effect; member states have not taken any measures to enforce compliance, and in many cases they have treated browser cookie settings as adequate implementation (see [32]).

In February 2012 the European Commission proposed a new set of revisions to EU data protection law [59]. Recommended provisions would alter the FTC’s present leadership on web tracking issues, and the Chairman of the FTC has suggested he shares that view [51].
Social networks and content providers are almost entirely absent. The DAA announced in late 2011 [63] that it would attempt to expand its program to non-advertising businesses and that it would broaden its consumer choice requirement to nearly all uses of third-party data for device personalization. Most of the largest social networks and content providers were stakeholders in the DAA’s program expansion and have not signed it.

There has been scant industry enforcement against busi-nesses that violate self-regulatory principles. In late 2011 the Better Business Bureau announced its first “decisions” against companies that had defective opt-out controls (see Section IX-A); the companies fixed their opt-out cookies, but were not otherwise penalized [66]. The NAI has released an annual “Compliance Report” since 2009 [67], [68], [69]. Only one company has been penalized for non-compliance; it is required to undergo an annual independent privacy audit for three years.

V. BUSINESS MODELS AND TRENDS

There are, broadly, six common business models for third-party websites: advertising companies, analytics services, social networks, content providers, front-end services, and hosting platforms. This taxonomy is intended to assist re-searchers in modeling third-party businesses; in practice, many services cut across business models, and new business models are frequently attempted.

- **Advertising Companies**
  
  While pricing models in online advertising converged by the early 2000s on a small set of auction algorithms (see [70], [71]), marketplace structures continue to evolve. There are three main models: direct buy, ad networks, and ad exchanges.

  1) Direct Buy: In the oldest model of online advertising, advertisers (and agencies) cut deals directly with first-party websites (“publishers”). This approach fell into disfavor for most websites in the late 1990s through 2000s, but remains the dominant model for search engine and social network advertising. Direct buy, has, of late, experienced a renaissance among content publishers owing to the development of “private advertising exchanges,” real-time advertising auctions run by publishers. Many implementations of direct buy advertising, especially search and social network advertising, do not load content from third-party websites, and therefore do not raise tracking privacy concerns.

  2) Advertising Networks: By the late 1990s growth in advertiser demand and ad slot supply (“inventory”) made it impractical for advertisers and publishers to deal directly. Ad networks offered a solution by enabling advertisers to easily place ads with many publishers, and by allowing publishers to support their content with many advertisers— with no ad sales team. Networks also brought the ability to systematically target ads to users, based on a publisher’s esti-mated audience (“demographic targeting”), a user’s location (“geographic targeting”), a web page’s content (“contextual targeting”), or a user’s browsing history (“behavioral target-ing”). Ad networks remain the largest and most widely used intermediaries in online advertising.

  3) Advertising Exchanges: In the mid-2000s publishers began seeking ways to monetize the “remnant” inventory they were not able to sell through an ad network. Ad exchanges offered to fill the slots, in real time, by enabling publishers to support their content with many advertisers—and with no ad sales team. Networks also brought the ability to systematically target ads to users, based on a publisher’s esti-mated audience (“demographic targeting”), a user’s location (“geographic targeting”), a web page’s content (“contextual targeting”), or a user’s browsing history (“behavioral target-ing”). Real-time bidding (or “RTB”). Ad exchanges quickly extended beyond remnants, and a number of intermediary business models now exist in the exchange ecosystem.

- **Content Providers**
  
  Content providers host video, maps, news, weather, stocks, and other media for embedding into websites. Some content providers, including YouTube, offer third-party wide-plug-ins to both increase user engagement and generate revenue through in-widget advertising. Many others, such as the Associated Press, charge for their content.

- **E Frontend Services**
  
  Some third parties provide services that assist publishers in distributing their own content, such as blog platforms (e.g. Wordpress.com) and content distribution networks (e.g. Akamai).

- **Market Trends**
  
  Krishnamurthy and Wills have collected longitudinal web measurements of approximately 1,200 popular websites be-tween 2006 and the present [74], [12], [13]. They report two consistent trends. First, tracking companies are rapidly increasing the share of websites that they span. Large trackers, including Google, Adobe, and Microsoft, have greatly extended their reach through acquisitions. Second, the number of trackers per page is growing rapidly. Websites now frequently embed content from dozens of third parties.

VI. ECONOMICS OF THIRD-PARTY WEB TRACKING

Proponents of web tracking often make the economic claim that it is needed to subsidize web services through advertising (e.g. [75], [76], [77], [78]). We believe the claim is subject to debate [79], and central questions remain open:

Which segments of the online advertising market de-pend on third-party tracking, and how is it used? It appears that only a small share of online advertising is behaviorally targeted [79]. The extent to which advertising relies on other uses of tracking is unclear.

What marginal tradeoffs do advertisers face for each use of tracking information? If tracking-based adver-tising becomes less feasible or more costly, advertisers will reallocate their expenditures. How they choose to reallocate will depend on the effectiveness and cost of the next-best alternatives to tracking-based advertising. Note that effectiveness and cost point in opposite directions—an advertiser may, for example, invest more in an advertising approach that is per-ad less effective but also per-ad significantly less expensive. To what extent can privacy-preserving technologies replace current uses of tracking? A number of designs have been advanced that, while not perfect substitutes, would enable much of the advertising functionality that tracking supports (see Section VIII-A). Limitations on tracking could incentivize advertising companies to de-develop and implement privacy-preserving technologies. What proportion of users would consent to tracking or pay if required to access a service? If diminished tracking-based advertising does impact publishers, they could require visitors to either pay or consent to track- ing. Some proportion of users would choose either option rather than forgo the service.
Given the public attention to third-party web tracking, there is surprisingly scant research on these central issues. A 2009 industry-sponsored paper by Beales [80] has been widely cited (e.g. [50]) for the proposition that behavioral targeting brings in substantially more value than other forms of ad targeting. Beales’s study focused on behaviorally targeted advertising which was roughly twice as expensive and twice as effective as untargeted (“run of network”) advertising. There are at least three problems with the methodology used in the study. First, the paper relies on data from a small, unrepresentative sample of advertising networks. Some statistics rely on data from fewer than five companies. The participating companies self-selected and were aware of the purpose of the study. Second, the paper compares behavioral advertising to untargeted advertising. As noted earlier, the relevant comparison is to the next-best alternative (e.g. contextual targeting). Third, the study concludes that behavioral advertising brings value to publishers through increased effectiveness and price. But, as noted earlier, increased price decreases the marginal value of behavioral advertising to advertisers.

Proponents of third-party web tracking have also frequently cited a 2011 paper by Goldfarb and Tucker [81], [82] reporting a 65% decrease in EU advertising effectiveness after the 2002 ePrivacy Directive was transposed by member states. We find four flaws in the Goldfarb and Tucker study. First, the analysis relies exclusively on self-reported data from one company’s surveys of web users. The paper does not explain how the data was collected, let alone demonstrate how it is valid and reliable. In fact, the survey data appears to have a number of oddities. It suggests, for example, that after the EU ePrivacy Directive non-EU advertising was twice as effective on EU viewers as on non-EU viewers. Second, the Goldfarb and Tucker data is not controlled for types of advertising. Behavioral advertising may only account for a slight share of the advertising in the study.

Third, the Goldfarb and Tucker study appears to incor-rectly assume that the 2002 EU ePrivacy Directive significantly altered online advertising behavior in Europe. In fact, advertising practices in the EU were largely unaffected by the ePrivacy Directive (see Section IV-C).

Fourth, the study seems to overlook changes in the online advertising market. Behavioral advertising was scarce in 2001 and a very small share of online advertising in 2008 [79]. The same time period yielded significant advances in contextual and demographic ad targeting. If the EU law negatively affected behavioral advertising, we should expect an across-the-board performance lift for EU and non-EU ads, with a slightly greater rise in non-EU performance. Instead, the authors predict and find four issues that might account for a slight share of the advertising in the study.

There are at least three problems with the methodology used in the study. First, the analysis relies exclusively on self-reported data from one company’s surveys of web users. The paper does not explain how the data was collected, let alone demonstrate how it is valid and reliable. In fact, the survey data appears to have a number of oddities. It suggests, for example, that after the EU ePrivacy Directive non-EU advertising was twice as effective on EU viewers as on non-EU viewers. Second, the Goldfarb and Tucker data is not controlled for types of advertising. Behavioral advertising may only account for a slight share of the advertising in the study.

The paper does not examine behavioral advertising in practice or third-party behavioral advertising.

**THIRD-PARTY WEB TRACKING TECHNOLOGY**

While the debates surrounding web tracking tend to focus on HTTP cookies, there are myriad stateful (“supercookies”) and stateless (“fingerprinting”) technologies that can be used to pseudonymously correlate web activities.49

<table>
<thead>
<tr>
<th>TABLE III</th>
<th>NON-COOKIE WEB TRACKING TECHNOLOGIES</th>
</tr>
</thead>
<tbody>
<tr>
<td>(a) “Supercookies”</td>
<td>HTTP authentication [84] HTTP caching (“cache cookies”)</td>
</tr>
<tr>
<td></td>
<td>Cache control</td>
</tr>
<tr>
<td></td>
<td>ETags (“ETag cookies”) [85] Last-Modified [85] (e.g. [96])</td>
</tr>
<tr>
<td></td>
<td>Content cache</td>
</tr>
<tr>
<td></td>
<td>Resource (e.g. JavaScript, HTML, CSS, or media) status code</td>
</tr>
<tr>
<td></td>
<td>Redirect location (e.g. [87]) hits and misses (e.g. [88])</td>
</tr>
<tr>
<td></td>
<td>TLS/SSL session ID [89]</td>
</tr>
<tr>
<td></td>
<td>browsing history</td>
</tr>
<tr>
<td></td>
<td>user Data storage (Internet Explorer only) HTML5 storage (session, local, and global) HTML5 protocol handlers</td>
</tr>
<tr>
<td></td>
<td>HTML5 content handlers1</td>
</tr>
<tr>
<td></td>
<td>W3C geolocation API permission2 window.name properties (session only) HTTP strict transport security [90]</td>
</tr>
<tr>
<td></td>
<td>Plug-in storage (e.g. Flash local shared objects, or “Flash cookies”) DNS cache</td>
</tr>
</tbody>
</table>

**A. Stateful Tracking (“Supercookies”)**

A website can encode a globally unique pseudonymous device identifier into any stateful web technology so long as it persists at least log-\(n\) bits, where \(n\) is the number of Internet-connected devices (presently roughly 5 billion, requiring 33 bits). Table III(a) provides a list of commonly deployed stateful web technologies and notes which have been observed in use for third-party web tracking. The evercookie library [92] provides a reference implementation for many of these tracking techniques.

Soltani et al. [14], McDonald and Cranor [93], and Ayenson et al. [8] report extensive use of Flash storage by popular websites, and Ayenson et al. found some use HTML5 local storage.

A number of online advertising companies, including ClearSpring, Interclick, Specific Media, and Quantcast, have been discovered using Flash cookies to track users. In mid-2011 Soltani [9] found that a third-party analytics service, KISSmetrics, was using cookies, Flash cookies, ETag cook-ies, cache cookies, userData, and HTML5 local storage, and that the non-cookie tracking technologies were used to recreate a cookie if deleted. We discovered that Microsoft was using an ETag cookie and a cache cookie in connection with its script for syncing an advertising identifier across web properties [94].

**B. Stateless Tracking (“Fingerprinting”)**

A website may be able to learn properties about the browser that, taken together, form a unique or nearly unique identifier [95], [96]. Some properties are passive, for instance, the active discovery of a script or plug-in in Table III(b)). Other properties can be passively learned from network traffic (Table III(c)).

In a 2010 sample of nearly 500,000 browsers Eckersley reported 83.6% were uniquely identified with a subset of active fingerprinting features. 94.2% of browsers with Flash or Java enabled were uniquely identified. While fingerprinting changed quickly, a simple matching algorithm was able to associate new and old fingerprints with over 99% precision.

Several companies, including 41st Parameter/AdTruth, BlueCava, and iovation, advertise commercial browser fin-gerprinting technology. Passive fingerprinting is particularly problematic since it cannot be detected with web measurement. Further research is needed to understand how effective passive fingerprinting is and what steps websites can take to scrub passive fingerprinting data from their logs. A recent study of Hotmail and Bing users by Yen et al. [97] suggests passive fingerprinting may be sufficient to track many stationary browsers.

**VIII. PRIVACY-PRESERVING THIRD-PARTY SERVICES**

There have been several efforts at designing third-party services that would capture the economic value of particular uses of tracking while preserving user privacy. Current proposals are based on a narrow subset of business models.

Further work is needed to support privacy across the range of evolving third-party website business models (Section V).

**A. Behavioral Advertising**

Privad [98] is designed to conceal a user’s activities from an advertising network by interposing an anonymity proxy between the browser and the ad network. In this approach, trusted client software subscribes to streams of possibly relevant ads, selects relevant ads locally,
submit candidates for auction, and then reports results. While the Privad model is designed to offer comprehensive privacy guarantees, it requires broad adoption of high-performance anonymizing proxies. This seems unlikely in the near future.

In an extension to Privad [99], Reznichenko et al. evaluate designs for privacy-preserving advertising auctions. The work emphasizes the tradeoff between an advertising company’s ability to conceal its ranking algorithm and bids and a user’s ability to prevent pseudonymous tracking.

Like Privad, Adnostic [100] uses client-based function-ality to perform ad selection, but it eliminates anonymizing proxies at the cost of less precision. Adnostic also simplifies the click-per-click billing model by allowing the advertising network to learn of a user’s ad clicks. Cost-per-impression billing would still require a low-performance trusted intermediary so as to not reveal the user’s ad impressions. As implemented, Adnostic requires a browser extension, which is a practical barrier to more widespread adoption.

RePriv [101], by Fredrikson and Livshits, is a verifiable policy architecture that enables users to selectively grant permission for generating and sharing client-side data stores that enable website personalization. The RePriv model holds promise as a general-purpose platform for building privacy-preserving advertising like Privad and Adnostic. But, like Adnostic, RePriv would have to be translated from its current implementation as a single-platform browser extension into existing web technologies for near-term deployment.

Bilenko and Richardson [102] propose an approach for keyword-based search advertising that provides privacy against a weaker threat model. The search advertising company is trusted to temporarily compute on user profile data, but then store the data in the browser and delete its copy. The authors ran their algorithm against 60 days of Bing search advertising logs and achieved almost all the benefit of cur-rent-server-side behavioral targeting. Specifically, they report capturing over 95% of the increase in click-through rates, generating approximately 4% greater revenue than search advertising without behavioral targeting. We are skeptical that the temporary data-use model is likely to be adopted; web services in general, and online advertising companies in particular, have historically been loath to voluntarily discard logs. The model also introduces the risk of inadvertent or surreptitious collection of third-party tracking data.

B. Analytics

Some analytics services have taken technical and legal precautions to silo tracking data for each first-party website.

Several free and paid services, including Google Analyt-ics and Adobe SiteCatalyst (formerly Omniture), use the same-origin policy to restrict the scope of pseudonymous identifiers to a first-party website. Google uses a first-party cookie to achieve this; Adobe offers the choice of a cookie scoped to a first-party subdomain CNAMEd to Adobe (e.g. metrics.apple.com) or a cookie scoped to a unique Adobe subdomain (e.g. pcpay.112.2o7.net).

Google Analytics offers an opt-in feature to websites that prevent the last octet of a user’s IP address (anonymized).26 This privacy option does not seem to reduce the benefit of the service since Google Analytics does not report IP addresses, and geolocation (the only reported measurement that relies on IP addresses) is unlikely to vary much by the last octet. We nonetheless found barely any use of the option: in an August 2011 crawl of the Alexa top 10,000 global websites, anonymized was set on only 63 of 4861 (1.3%) reports to Google Analytics.

Paid analytics services usually promise by contract to make no use of the data they collect except as directed by their clients, and they impose internal business controls to ensure each client’s data remains segregated. Adobe, for example, makes these guarantees [103]: “Although the data generated by Adobe’s products resides on Adobe’s servers, each customer owns the data generated by the use of its site. By contract, Adobe has no right to access or use this data. In addition, Adobe does not allow use of the data for any purpose other than those of the owner (web publisher), that is, Adobe silos each customer’s data for use by that customer.”

IX. USER CHOICE MECHANISMS

Three technical solutions have been advanced for giving users control over third-party web tracking: opt-out cookies, blocking, and Do Not Track.

A. Opt-Out Cookies and the AdChoices Icon

User choice in current online advertising self-regulation is implemented with opt-out cookies. There are several problems with this approach. First, and less concerning, is that it is not adequately updated. To opt out of new third parties, a user has to install new cookies. Second, cookies expire, so a user has to periodically renew opt-out cookies. Third, users may clear their cookies, inadvertently removing their opt-out preferences. Fourth, opt-out cookies are fragile; it is easy for a third party to improperly set or delete an opt-out cookie. Fifth, opt-out cookies scale poorly; each third-party PS+1 require a network roundtrip, resulting in a sluggish user experience when changing multiple preferences. Browser extensions for persisting opt-out cookies, such as TACO or Google Keep My Opt Outs, largely mitigate these issues at the cost of usability.

Many online advertising companies have begun to insert an “AdChoices” icon (13x13px) and text (10pt) into display ads (Figure 2(b)) to increase user awareness of behavioral targeting and existing self-regulatory mechanisms. Clicking the icon provides additional information about how the ad was targeted and, in many cases, a link to landing page where the user can set opt-out cookies.

Several studies have called into question the usability of the self-regulatory opt-out model.

(a) Proposed icon and text [104] (actual size at 115 DPI).

(b) Implemented icon and text [105] (actual size at 115 DPI).

Before the deployment of the AdChoices icon an industry-funded policy group conducted a large-audience usability survey [106]. It found that a 31x31px icon with 18pt font (Figure 3(a)) was not very effective at conveying information about behavioral targeting practices (“substantial repetition and consumer education may be needed to improve the icon’s communication effectiveness over time”), and that the text “AdChoice” performed worse than alternatives.

McDonald and Conran [34] conducted a large-audience survey on user perceptions of a self-regulatory opt-out page. 88% of participants understood that the page related to online advertising and opting out, but only 11% correctly responded that the page allows opting out of behavioral targeting, not tracking (34%), ads from specific companies (25%), or some proportion of advertising (18%).

Leon et al. [107] examined the usability of two other self-regulatory websites with five in-laboratory participants each. On one website, the Digital Advertising Alliance, only one of the five users was able to opt out without guidance, and none of the users correctly understood the implications of opting out. On the other website, Evidon, four of the five users were able to opt out without guidance, though it took the participant who chose to opt out of all companies 47 minutes to exercise his or her preferences. Once again none of the users correctly understood what opting out would do.

Leon et al. also studied the usability of the TACO exten-sion. All five participants enabled persistent opt-out cookies, the default setting.

Hernandez et al. [105] measured the prevalence of the AdChoices program on the Alexa U.S. 500 top homepages. They found an icon in only 9.9% of third-party ads and an icon and text in only 5.1%.

The online advertising trade groups have declined to provide overall usage statistics of opt-out cookies. Anecdotal reports (e.g. [108]) place usage at less than 1% of browsers.

B. Blocking

Given the myriad approaches to tracking a browser many of which require nothing more than an HTTP round trip—the most effective user self-help tools27 function by blocking third-party web content. Nearly all tools consist of a block list, either available as a subscription for a browser extension or wrapped in a configurable browser extension.

To understand the effectiveness of blocking, we conducted three consecutive FourthParty crawls of the Alexa U.S. top 500 with each of 11 blocking tools installed [110]. We also conducted a baseline crawl to estimate which PS+1 were third-party trackers. For each tool we calculated three values relative to baseline and averaged across all trackers: pages with an HTTP request to a tracker, pages with an HTTP Set-Cookie response from a tracker, and cookies added less cookies deleted by a tracker.28 Figure 3 presents our results.

We found significant variability in performance. The most effective tool was a combination of community-maintained Fanboy’s Lists for blocking ads, surreptitious tracking, and social content. All of the top

www.ijcsit.com 7485
performing tools blocked third-party advertising, an unsurprising result since there is no clear division between advertising content and advertising-related tracking content. The block list from TRUSTe was not only the least effective, but it also would override other lists to allow tracking by several sizable third parties.\(^1\)\(^2\)\(^3\)\(^4\)\(^5\)

Leon et al. [107] examined the usability of the Ghostery, AdBlock Plus, and Internet Explorer Tracking Protection List blocking tools. Two of the five Ghostery users believed they had enabled the extension’s blocking feature when, in fact, they had not. All five of the AdBlock Plus users configured the extension with a default advertisement blocking list; none installed additional lists to block non-advertising track-ers. All five Internet Explorer Tracking Protection List users retained the default setting, to not block any content; they all believed they had configured the option to substantially or completely block tracking.

In sum: blocking can be fairly effective, but it is only a realistic solution for advanced users.

C. Do Not Track

Do Not Track uses a combination of technology and policy to provide consumer choice over web tracking. The World Wide Web Consortium (W3C) is presently standard-izing Do Not Track; the W3C’s working group has not yet reached consensus on the technology or policy components.

The Do Not Track technology is simply an HTTP header, DNT, that signals a user’s preference about web tracking. As-sociated technologies have been proposed that would allow a website to request exceptions and signal its own tracking status. Firefox, Internet Explorer, Safari, and Opera presently support a Do Not Track opt-out preference (sending the DNT: 1 header). Google has pledged to add the feature to Chrome. As of late 2011, Mozilla [111] reported 5.6% usage in desktop Firefox and 17.1% usage in Firefox Mobile.

Roughly twenty websites presently honour the Do Not Track technology, and the Digital Advertising Alliance recently pledged [39] that its about eighty member companies would begin supporting the header.

Do Not Track enforcement could be accomplished through measurement of tracking technologies, using tools like FourthParty.\(^\cdash7\) In mid-2011 we identified two advertising companies that were surreptitiously taking steps to honour Do Not Track [113], suggesting the approach is quite viable.

The Do Not Track policy defines what websites must do when they receive a Do Not Track header. Debates over the Do Not Track policy have been largely coextensive with debates over third-party web tracking policy (see Section III-E). Policymakers, consumer advocates, and researchers are in general agreement that Do Not Track must significantly curtail third-party information collection. The recent DAA commitment was largely coextensive with debates over third-party web tracking as of early 2012. The field is rapidly changing; new tracking by several sizable third parties.

In conclusion, this paper surveyed policy and technology issues in third-party web tracking as of early 2012. The field is rapidly changing; new announcements, questions, and research re-sults appear by the week. We hope the information presented here provides security and privacy researchers with the background necessary to contribute to this developing field and to meaningfully participate in the on going public debate.

REFERENCES


[33] A. Narayanan and L. F. Cranor, “Informed users’ understanding of behavioral advertising,” in Proceedings of the 2010 Research Conference on Com-merce-ization and Information and Internet Policy,


Confidentiality, 2011.


M. Janssen, "Do Not Track" rules would put a stop to the Internet as we know it," U.S. News, January 2011, opinion.


J. Hodges, C. Jackson, and A. Barth, "HTTP transport security (HSTS)," draft-ietf-websec-strict-transport-ses-06, March 2012.


