Web privacy measurement in real-time bidding systems. A graph-based approach to RTB system classification

Eijk, R.J. van

2

Literature review

In this chapter, we provide a literature review. It is the third step of our research methodology (Section 1.5). We start by giving an overview of Web Privacy Measurement (WPM) (Section 2.1). This is followed by three client identification mechanisms (Section 2.2), deterministic and probabilistic identification (Section 2.3), web-based fingerprinting (Section 2.4), HTTP header fields (Section 2.5), an overview of four different approaches to data collection (Section 2.6) in the academic field of WPM, and an overview of ten WPM visualization tools (Section 2.7). Section 2.8 concludes the chapter.

2.1 Web Privacy Measurement

In this thesis we adopt the definition of Web Privacy Measurement as proposed by Englehardt, Eubank, Zimmerman, Reisman, and Narayanan [2014].

Definition 2.1: „Web Privacy Measurement is the observation of websites and services to detect, characterize, and quantify privacy-impacting behaviors." [Englehardt et al., 2014]

WPM has been established as an academic research field in 2012.[66] The key objective of the research field is to increase transparency through measurement. Several components including the identification of the (third-party) web tracking problem, were discussed

66 We mark the launching of the research field by the organization of the workshop web privacy measurement hosted by the Berkeley Center for Law & Technology, in 2012.

during the workshop web privacy measurement hosted by the Berkeley Center for Law & Technology (see, e. g., McDonald and Van Eijk [2012]). It is situated in the intersection of computer science and law.[67]

One of the foremost outcomes of the workshop web privacy measurement hosted by the Berkeley Center for Law & Technology was a literature review composed by Krishnamurthy and Wills [2012].[68] The review contained 81 publications. The literature was grouped in seven dominant WPM themes, i. e.,

(1) characterization of tracking browser behavior,
(2) leakage of (personal) information to third parties,
(3) extraction of (personal) information from collected data,
(4) linking (personal) information,
(5) privacy protection and privacy preservation tools,[69]
(6) privacy and economics, and
(7) end-user attitudes regarding privacy.

The outcome led to Hoofnagle and Good [2012b], which contained an overview of WPM studies with a focus on „discovering tracking vectors and quantifying them." Both reviews put WPM in a historical context. Most notably we refer to Krishnamurthy and Wills [2006], Krishnamurthy, Malandrino, and Wills [2007], Krishnamurthy and Wills [2009a; 2009b], and Gomez et al. [2009] for their ground-breaking work.

We consider the two literature reviews mentioned above as a point of reference for our efforts to provide new insights into WPM. Next, we provide a brief overview of why web privacy measurement matters (Subsection 2.1.1), event tracking (Subsection 2.1.2), give an example of an action, viz. Google Analytics (Subsection

67 See, e. g., Andersdotter, Castex, Dubois, Hamilton, Madelin, and Van Eijk [2014]; Jayaram, Kutterer, Kwasny, Runnegar, Seltzer, Van Eijk, and Zorbas [2012].

68 The review is not easy to find on the internet. It was posted in an email with the subject 'WPM Privacy Measurement Bibliography follow-up' to the (non-public) conference mailing list [Wills, 2012]: „Fellow WPMers - Following up on the conference, Bala [Balachander Krishnamurthy] and I finally got around to updating the Web Privacy Measurement Bibliography that we distributed at the conference. Based on feedback, we added a dozen additional entries. You can find the updated files at: `http://web.cs.wpi.edu/~cew/share/wpm.pdf` (27 December 2015) and `http://web.cs.wpi.edu/~cew/share/wpm.bib` (27 December 2015).
Apologies if we did not include something that was sent, but we tried to stay with the focus on web privacy measurement. - Craig Wills."

69 Van Eijk [2011b] is listed under the theme privacy protection and privacy preservation tools.

2.1.3), discuss information asymmetry (Subsection 2.1.4) and introduce a new multidisciplinary legal-science paradigm (Subsection 2.1.5). We remark that the literature reviewed in Sections 2.2–2.6 relates to RQ1 (Chapter 3). For RQ2 (Chapter 4) we will explore new territory, i. e., the application of network science algorithms to WPM. To anchor our exploration, we provide a brief overview of other graph-based approaches in Section 2.7.

2.1.1 *Why web privacy measurement matters*

The Snowden revelations and the Schrems-Facebook cases have shown the significance of the collection of *metadata* and the collection of *communications*. Section 215 of the US Patriot Act contains the legal basis to collect *metadata* of phone calls made in the US.[70] Section 702 of the Foreign Intelligence Surveillance Act (FISA) contains procedures for the collection of *communication* of certain persons outside the United States other than United States persons.[71]

We refer specifically to the memo by Farrell and Tschofenig [2014, RFC 7258],[72] the reports by Farrell, Wenning, Bos, Blanchet, and Tschofenig [2015, RFC 7687],[73] and Barnes, Schneier, Jennings, Hardie, Trammell, Huitema, and Borkmann [2015, RFC 7624],[74] but also to the publications by, e. g., Gallagher and Greenwald [2014], Nottingham [2015], Poitras [2014], Soltani [2014], Snowden [2014], Hilts and Parsons [2015], Marquis-Boire, Greenwald, and Lee [2015], Mayer [2015a], Swire [2015], and Swire [2017].

The Snowden revelations clearly confirmed two things:

(1) the mechanisms as identified by, e. g., Gomez et al. [2009], Mayer and Mitchell [2012], and Janc and Zalewski [2014] have very practical applications, and

70 Swire [2017, p. 1-11]: „Perhaps the most dramatic change in US surveillance statutes since 2013 concerns reforms of Section 215 of the US Patriot Act, which provided the government with broad powers to obtain 'documents and other tangible things.' 31 After the September 11 attacks, Section 215 was used as a basis for collecting metadata on large numbers of phone calls made in the US."

71 Swire [2017, p. 1-11]: „Section 702 of FISA applies to collections that take place within the US, and only authorizes access to the communications of targeted individuals, for listed foreign intelligence purposes.", Viz. 50 United States Court section 1881a URL: `http://uscode.house.gov/view.xhtml?path=/prelim@title50/chapter36&edition=prelim` (18 December 2015).

72 Title: Pervasive Monitoring Is an Attack.

73 Title: Report from the Strengthening the Internet (STRINT) Workshop.

74 Title: Confidentiality in the Face of Pervasive Surveillance: A Threat Model and Problem Statement.

(2) persistent (unique) identifiers (UIDs) enable identification of end-users when they use computers and smart mobile devices to access the web.

To illustrate the interconnectedness between State tracking and commercial tracking, we refer to Kaye [2015] - the Special Rapporteur on the promotion and protection of the right to freedom of opinion and expression. Kaye [2015] concludes that encryption and anonymity enable individuals to exercise their human rights and, as such, deserve strong protection. His predecessor La Rue [2011] already underlined the urgent need to further study new modalities of surveillance. Although both rapporteurs referred primarily to State surveillance, the argument in our opinion also holds for commercial tracking. Commercial tracking endangers a number of basic rights which, in aggregate form, constitute the foundation of liberal societies.[75] Based on the above, we remark that the tracking of data collected commercially is at the heart of WPM.

More recently, we witnessed a use of RTB for *political advertising online* (see, e. g., Lambiotte and Kosinski [2014], Youyou, Kosinski, and Stillwell [2015], or Berners-Lee [2017] who underlined the need for transparency and understanding in political advertising online).[76] We remark that the Interactive Advertising Bureau (IABureau) already addressed microtargeted political advertising in Election 2012:

(1) to recruit,
(2) to raise money,
(3) to persuade undecided voters, and
(4) to get out the vote (cf. Abse [2012]).[77]

75 In an interview Swire explained that the internet is much more than just a surveillance conduit: „One of our biggest themes [of the report ([Swire et al., 2013]) of The President's Review Group on Intelligence and Communications Technologies] was that these decisions about surveillance were not just the intelligence community's decisions, they were affecting our allies, our economic activities in the world, privacy and civil liberties." [Mayer, 2015a, 17"17′–18"19′].

76 On 5 July 2018, the European Parliament adopted a resolution to underline the importance to protect the European fundamental right to data protection and to ensure consumer trust. We remark that the EU-US Privacy Shield (approved in 2016) is the successor to the Safe Harbour framework (approved in 2000). URL: http://www.europarl.europa.eu/news/en/press-room/20180628IPR06836/suspend-eu-us-data-exchange-deal-unless-us-complies-by-1-september-say-meps (6 July 2018).

77 Abse [2012]: „Microtargeted political ads are being used at all key points in political campaigns to recruit and raise money, to persuade undecided voters and

Kosinski [2014] showed the effectiveness of personality targeting by matching products and marketing messages to end-users' personality characteristics. The characteristics are derived from a model based on „dimensionality reduction for preprocessing Facebook-likes data, which are then entered into logistic/linear regression to predict individual *psychodemographic profiles* from likes" [Kosinski, Stillwell, & Graepel, 2013]. Moreover, Kosinski [2014] takes the social network structure of end-users into account and shows that differences in personality affect the online as well as the offline behavior of individuals and groups.

In an attempt to address the risks of political advertising online Chester and Montgomery [2017] emphasized two issues: (a) they examined the application of *personality targeting* during the most recent election cycle in the United States and (b) explored the implications of their continued use.[78]

2.1.2 *Event tracking*

New business models and the technical abilities to track people when they are online must coincide to become a business success (i. e., to mature the own development). For instance, standard page tracking is moving to a new level: event tracking.[79] We start by giving the definitions we use in our research.

DEFINITION 2.2: EVENT TRACKING is a method for recording in *real time* the interactions that visitors have with website elements.

The key objective of event tracking is to measure all the interactions that people have *when* they engage with online content and are trying to understand the data that comes with it. The data analytics applied to the visitor gives meaning to the events.[80]

---

to get out the vote. They make use of online and offline data to find appropriate audiences, and create constantly-adjusted models to further refine their focus."

78 Personality targeting is a type of behavioral targeting. Cambridge Analytica used a model known as OCEAN, which processes data from individuals supplied by „leading companies, including Axciom, Experian, Nielsen, GOP firm Data Trust, Aristotle, L2, Infogroup, and Facebook." [Chester & Montgomery, 2017, p. 7]

79 Cf. Rappaport [2015, p. 144]. See also Ellis [2014]; Thibault [2013].

80 See Definition 1.10 on data.

In fact, display advertising relies on a complex network of interdependent business processes.[81] The network consists of four *steps*.[82]

STEP 1. Building the ad campaign to reach an audience.
STEP 2. Buying advertising in terms of frequency and reach of the audience.[83]
STEP 3. Tracking and optimizing the campaign depending on the campaign goals.
STEP 4. Reporting on the campaign results (campaign analytics) and audience results (audience analytics).[84]

Basically, event tracking guides marketers by moving people through four distinct marketing *stages*.[85]

STAGE 1. Catching the awareness of the end-user.
STAGE 2. Getting in touch with the end-user.
STAGE 3. Making sure the value proposition leads to action.
STAGE 4. Continue the customer relation for future interaction.

Nowadays most large players in the digital marketing landscape offer analytics packages capable of event tracking. Below, we provide five examples: (A) Google, (B) Yahoo, (C) Twitter, (D) Facebook, and (E) Oracle.

A: GOOGLE collects event-tracking data „to help and analyze visitor traffic and paint a complete picture of the audience and their needs, wherever they are along the path to purchase."[86]

81 See note 25. See also Figure 1.1.
82 Cf. Outsource Ad Ops [2014; 2015]; Reagan [2013], see also Kotler and Armstrong [2012, p. 437].
83 Four common ways of pricing an ad campaign are Cost-Per-Mille (CPM), Cost-Per-Click (CPC), Click Through Rate (CTR), and Cost Per Acquisition (CPA). CPM is based on impressions of the ad, the remuneration is fixed on a price per 1,000 views. CPC is based on clicks on the ad by the intended audience, the remuneration is fixed on a price per click. CTR measures the efficiency of the ads as a percentage of the clicks per ad impression. CPA is based on the intended audience buying an advertised product or service, the remuneration is fixed on a price per completed sale. (cf. AppNexus [2017c]).
84 Whereas campaign analytics give insight into how a campaign performs, audience analytics give insight into how an audience engages with a campaign through the seven different internet platforms (Figure 1.1).
85 Cf. [Rappaport, 2015, pp. 12–13,40–52] used the following terms: (1) capture, (2) connect, (3) close, (4) keep.
86 URL: `http://www.google.nl/analytics/why/` (1 February 2015).

B: YAHOO collects event-tracking data with web beacon technology and uses it to serve personalized advertisements to end-users based on their interactions with a website.[87]

C: TWITTER collects event-tracking data of followers to explore their interests, locations, and demographics.[88]

D: FACEBOOK collects event-tracking data to understand who uses apps, how they use the apps, optimize their experience, and reach them with powerful ad campaigns. The data provides audience insights, e. g., age, gender, education, interests, country, language.[89]

E: ORACLE collects event-tracking data in real time for various purposes, e. g., churn prediction, product recommendations, and fraud alerting.[90] Churn is profiling aimed at identifying customers that a company is about to lose.

In summary, the five products noted above enable the measurement of unique end-users through all kinds of tracking events, across devices, whenever and wherever they connect to the web. The products use unique identifiers for multiple purposes, e. g.,

(1) counting the number of times that an end-user sees a particular advertisement (henceforth: frequency capping),

87 The original description is as follows. „Yahoo Web Analytics uses web beacons to collect information on how visitors use our customers' websites, such as the types of pages viewed or interactions on those websites. By using this information, Yahoo delivers site analytics to its customers, which allows them to bring more relevant advertising, products and services to you. Yahoo also uses anonymous information gathered through web beacons to serve you more relevant advertising and to improve its products and services." URL: `https://reports.web.analytics.yahoo.com/optout,OptOut.vm` (1 February 2015).

88 URL: `https://analytics.twitter.com/about` (10 August 2016)

89 The original quote is as follows: „Facebook Analytics for Apps is a powerful solution to understand the people who use your apps, optimize their experience and reach them with powerful campaigns. With over a billion people on Facebook, Analytics for Apps is the only solution that gives you audience insights like age, gender, education, interests, country, language and many more. It's all free and you don't need to use Facebook Login." URL: `https://developers.facebook.com/products/analytics` (10 August 2016)

90 The original description is as follows. „With Oracle Advanced Analytics, customers have a comprehensive platform for real-time analytics that delivers insight into key business subjects such as churn prediction, product recommendations, and fraud alerting." URL: `http://www.oracle.com/us/products/database/options/advanced-analytics/overview/index.html` (6 February 2015).

(2) ad verification to prevent billing advertisements not shown to unique visitors,
(3) to randomly assigns the end-user to version A or B of an object and measure their response (henceforth: A/B testing),
(4) security/fraud detection, and
(5) measurement of conversion rates.[91]

2.1.3 *Example of an action: Google Analytics*

To provide somewhat more insights into the consequences of the descriptions given above, we elaborate on example (A) Google Analytics.

In fact, Google's analytics measurement protocol enables developers to make HTTP requests to send raw end-user interaction data *directly* to Google Analytics servers.[92] Google explains to developers that they can:

(A) measure detailed end-user activity,
(B) transfer online data to offline data, and
(C) send data from both the browser and the server.

Furthermore, developers can send data *indirectly* to Google Analytics servers, e. g., by using Google's data layer [Google, 2017a; 2017b]. The data layer provides the developer with a mechanism to indicate, e. g., that an end-user has been identified as being a prospect with buying intentions.[93]

However, the information provided to website owners is quite different from the information provided to the end-user. For ex-

91 Cf. IABureau [2017c], Doty, West, Brookman, Harvey, and Newland [2016, Subsection 3.3.2 Permitted Uses], Audit Bureau of Circulations [2015], and Ivie, Gunzerath, and Pinelli [2015], viz. Definition 1.8. Current fraud related challenges RTB is facing are: (1) domain fraud „to change the domain to a more valuable publisher", (2) location fraud „to change the IP address or location to a more valuable location", (3) user Identifier (ID) fraud „to change the device ID or buyer id to an ID that has historically monetized well", and (4) device fraud „to change from a type of device that doesn't tend to monetize well, e. g., as a smartphone, to one that does, e. g., a desktop" [IABureau, 2017b, p. 2].

92 Google Analytics Measurement Protocol developer's guide: „The Google Analytics Measurement Protocol allows developers to make HTTP requests to send raw end-user interaction data directly to Google Analytics servers. This allows developers to measure how end-users interact with their business from almost any environment." URL: `https://developers.google.com/analytics/devguides/collection/protocol/v1/` (20 July 2015).

93 I am indebted to Towvim [2018] for this insight.

ample, on 13 July 2015 Google provided the following prominent privacy reminder to its Dutch end-users: „Website owners use data collected with Google Analytics to improve the operation of their websites."[94] Obviously Google leaves it up to a website owner to further explain what the term improvement means when an end-user visits its website.

2.1.4 *Information asymmetry*

The consequence of the proliferation of event tracking is the appearance of information asymmetry. The aim of WPM is to mitigate the information asymmetry caused by contemporary digital data collection practices. For instance, Englehardt, Acar, and Narayanan [2017] reported on the use of event-tracking scripts in webpages collecting event data about end-user's page interactions, e. g., keystrokes, mouse movements, and scrolling behavior.[95] Unlike typical analytics services that provide aggregate statistics, these scripts are intended for the recording and playback of individual browsing sessions, as if someone is looking over your shoulder

The asymmetry consists mainly of two elements: (1) inadequate (real-time) information provided to the end-user, and (2) unclear privacy policies. Both are essential for adequate end-user empowerment and end-user trust in online services. A variety of scholars have addressed the asymmetry, see, e. g., Levin and McCain [2014], Zuiderveen Borgesius [2014], Rich [2015], Thode, Griesbaum, and Mandl [2015], and Kulyk, Hilt, Gerber, and Volkamer [2018].

2.1.5 *Multidisciplinary legal-science paradigm*

Ermakova, Fabian, Bender, and Klimek [2018] performed a recent literature review on the state of research in WPM.[96] They catego-

94 Screenshots of the privacy reminder were retained on 13 July 2015. The original Dutch message was as follows. „Website-eigenaren gebruiken gegevens die worden verzameld door Google Analytics om hun sites beter te laten werken. U kunt zich voor deze gegevensverzameling echter afmelden door een add-on voor uw browser te downloaden en te installeren."

95 Englehardt et al. [2017] denotes event-tracking scripts as session replay scripts.

96 They followed the five-step approach by Herz, Hamel, Uebernickel, and Brenner [2010, pp. 2–9]: (1) definition of review scope, (2) conceptualization of the topic, (3) literature search, (4) literature analysis and synthesis, and (5) research agenda.

rized 31 WPM publications into six research methodologies (see Table 2.1, denoted by the first column 'Research methodology'). Furthermore, Ermakova et al. [2018] grouped the research methodologies into two paradigms, i. e.,

(1) design-science paradigm (Table 2.1, second column) and
(2) behavioral-science paradigm (Table 2.1, third column).

Table 2.1: Overview of six WPM-research methodologies.

RESEARCH METHODOLOGY	DESIGN SCIENCE PARADIGM	BEHAVIORAL SCIENCE PARADIGM	MULTI-DISCIPLINARY LEGAL SCIENCE PARADIGM
Argumentative-deductive analysis	✓		✓
Prototyping	✓		
Modeling	✓		✓
Qualitative-empirical cross-sectional analysis		✓	
Grounded theory		✓	
Field study		✓	

In the *design-science* paradigm, the understanding of a problem and its solution are achieved in building, e. g., models, methods, or systems [Wilde & Hess, 2007] (as cited by Ermakova et al. [2018]). In the *behavioral-science* paradigm theories for explaining or predicting behavior of individuals or organizations are central [Hevner, March, Park, & Ram, 2008] (as cited by Ermakova et al. [2018]).[97]

Earlier Geer [2015] has put forward the question (here paraphrased as): „Is the current paradigm of the privacy science already sufficiently in a crisis that a resolution of the crisis requires

97 Ermakova et al. [2018]: „The behavioral-science paradigm attempts to form and justify theories for explaining or predicting behavior of individuals or organizations [Hevner et al., 2008], whereas the design-science paradigm deals with developing and assessing IT artifacts (e.g., models, methods or systems) to enlarge their capabilities [Hevner et al., 2008; Wilde & Hess, 2007]."

a change of paradigm?"[98] I personally believe (and with me many others) that (1) the changes in the legal landscape for online privacy in Europe and (2) the advances in online advertising are currently sufficient causes of a crisis to require a formulation of a third paradigm of WPM.[99] Based on the above, I suggest to include the *multidisciplinary legal-science* paradigm (Table 2.1, fourth column). Two examples which we categorized as early WPM results in the multidisciplinary legal-science paradigm are Leenes and Kosta [2015] and Trevisan, Traverso, Metwalley, and Mellia [2017]. Both papers investigate the consent requirement for tracking cookies with argumentative-deductive analysis and WPM techniques. Furthermore, we refer to Turcios Rodríguez [2018, p. 16], Degeling, Utz, Lentzsch, Hosseini, Schaub, and Holz [2018], and Iordanou, Smaragdakis, Poese, and Laoutaris [2018].

We remark that the research methodology (see Section 1.5) is fit and tuned to include our novel graph-based approach as an example of the multidisciplinary legal science paradigm (more precisely, they are Step 5, viz. the development of a metadata model and storing the data into a graph and Step 6, being the analysis of the research data with intelligent techniques). Furthermore, we see that 'Argumentative-deductive analysis' is also fit and tuned for the new paradigm because the new paradigm aims to investigate legal issues through empirical exploratory research (Step 7, compilation of a normative framework and Step 8, the evaluation of the framework).

Obviously, new marketing and advertising techniques are currently collecting data for their actions in real time. In our research, we are investigating these actions. In Subsection 2.1.1 we focused on why web privacy measurement matters.

At this moment, we remark that the literature reviewed in this chapter relates to RQ1 (Chapter 3). For RQ2 (Chapter 4) we will explore new territory, i. e., the application of network science algorithms to WPM.

98 Geer [2015]: „Is the paradigm of privacy science in enough of a crisis that resolution of the crisis requires a change of paradigm?"

99 K. Thomas [1962, p. 103]: a paradigm is established if the following question is answered affirmatively: „are the source of the methods, problem-field, and standards of solution accepted by any mature scientific community at any given time?"

2.2 THREE CLIENT IDENTIFICATION MECHANISMS

In this section we aim to gain an insight into three of the latest client identification mechanisms. For the reader it is necessary to know that a UID may be communicated in various ways, e. g., (1) by a URL parameter, (2) by a HTTP POST-method, and/or (3) by a HTTP header-field. Let us therefore turn to two publications to guide our thinking. They are separately discussed in subsections 2.2.1 and 2.2.2.

2.2.1 *Technical analysis*

As one of the first, Janc and Zalewski [2014] performed a detailed technical analysis of client identification mechanisms. They rightfully point out that the current browser privacy controls evolved almost exclusively around the notion of HTTP cookies. They used three categories of client identification mechanisms:

(1) explicitly assigned client-side identifiers (UIDs),
(2) inherent client device characteristics that identify a particular machine, and
(3) measurable end-user behaviors and preferences that may reveal their identity.

Moreover, they referred to web tracking as „a process of calculating or assigning unique and reasonably stable identifiers to each browser that visits a website." The qualification 'reasonably stable' is interesting, because from a legal point of view the discussion whether a reasonably stable identifier is personal data puts indirect identifiability square in its center.

2.2.2 *Five categories of tracking mechanisms*

As a clear follow-up, Bujlow, Carela-Español, Solé-Pareta, and Barlet-Ros [2015] composed a timeline (1994–2014) which provides an overview of web tracking mechanisms based on their first occurrence.[100] The timeline [Bujlow et al., 2015, p. 2] overlaps with the analysis of Janc and Zalewski [2014]. However, it is far less detailed.

100 See also Bujlow, Carela-Español, Solé-Pareta, and Barlet-Ros [2017].

The main take away of the timeline is a description [Bujlow et al., 2015, pp. 3–13] of 31 tracking mechanisms grouped in five categories:

(1) session-only,[101]
(2) storage based,[102]
(3) cache based,[103]
(4) fingerprinting, and
(5) other web tracking mechanisms.

The first three categories have as their main goal the identification of the end-user. The identification can be done in two main ways: deterministic and probabilistic. We discuss the first three categories along these two ways in Section 2.3. Subsequently, we discuss web-based fingerprinting in Section 2.4. In the thesis we do not discuss other web tracking mechanisms, such as timing attacks,[104] clickjacking,[105] and evercookies (supercookies).[106] We only mention them for completeness.

2.3 DETERMINISTIC AND PROBABILISTIC IDENTIFICATION

Deterministic and probabilistic identification is the explicit assignment of online identifiers to end-users. Janc and Zalewski [2014] already gave us a picture of three client identification mechanisms. The first client identification mechanism is the explicit assignment of client-side identifiers to collect event data from end-users. Here, we would like to emphasize the two latest client identification methods: (1) deterministic methods and (2) probabilistic methods.[107] Client identification is basically the extraction of a unique

101 Websites use session storage to store IDs.
102 For instance, local storage, web SQL, HTTP cookies and the recent W3C Recommendation IndexedDB [Alabbas & Bell, 2018].
103 We mention cache storage and application cache, e. g., the recent Service Worker cache Application Programming Interface (API) [Russell, Son, Archibald, & Kruisselbrink, 2018].
104 Bujlow et al. [2015, p. 4]: „Boolean values dependent on the look of the website (e.g., if the user is logged in to a particular service), stealing any graphics embedded or rendered on the screen."
105 Bujlow et al. [2015, p. 4]: „User's email and other private data, Paypal credentials, spying on a user by a webcam."
106 Bujlow et al. [2015, p. 4]: „Operating system instance id, browser instance id." See also, e. g., Acar, Eubank, Englehardt, Juarez, Narayanan, and Diaz [2014]
107 See also 'The all-pervasive impact of Big Data in daily life: from Causality to Correlation' [Van den Herik & Van Eijk, 2014].

identifier. To differentiate between the two terms and simultaneously take the latest technological capabilities of the RTB landscape into account, we propose a definition for each method (Definition 2.3 and Definition 2.4), (see Subsections 2.3.1 and 2.3.2).

2.3.1 *Deterministic identification*

DEFINITION 2.3: DETERMINISTIC CLIENT IDENTIFICATION is a process of client identification with metadata directly linked to an individual.

Two examples of deterministic client identifiers are (relatively) stable UIDs such as, e. g., an end-user's email address or a device identifier of an end-user's smart mobile device. Deterministic client identifiers enable linking end-user behavior (1) across his devices, e. g., desktop browser or app and (2) across marketing channels, e. g., e-mail, call center, or social-network websites.

A recent example of a deployment of deterministic client identifiers to track the behavior of its customers is taken from KLM Royal Dutch Airlines (Van Eijk [2018, p. 10], Google [2017c]).[108] The KLM case study serves as an example of deterministic identification. KLM partnered with Google and Relay42 to collect „all customer interactions on KLM's website and app, as well as [the] relevant indicators from other channels and [other] data sources." The case study is exemplary for cross-device tracking by Tag-Based Integration (see Definition 4.11). In summary, the identifiers are deployed with Relay42 tags and DoubleClick Floodlight tags so that KLM is able to send the end-user metadata from Relay42's Data Management Platform (DMP) to the Google DoubleClick RTB system (cf. Google [2017c]).[109]

2.3.2 *Probabilistic identification*

DEFINITION 2.4: PROBABILISTIC CLIENT IDENTIFICATION is a process of client identification based on UIDs derived from metadata relating to an individual and based on probability.

108 Infra Subsection 4.3.8.

109 The full quotation is as follows: „Data gathered via Relay42's own tags and DoubleClick Floodlight tags can also be sent from the DMP to the DoubleClick platform so that relevant ads can be targeted to appropriate audiences."

The KLM case study [Google, 2017c] also serves as an example of probabilistic identification.[110] To correlate site behavior with ad impressions, KLM stores their cross-device data in Google BigQuery [Grigorik, 2013]. In this way, KLM can correlate site behavior with targeted ads. Relay42 has a key role in this process, since they add metadata to KLM's dataset and build predictive models. Google illustrated the new capabilities as follows.

> „KLM developed a real-time buying setup to include predictive modeling. In this setup, website and app interactions are tracked in the DMP thanks to the Relay42 tag management system. Relevant consumer behaviour can be streamed in real-time to a prediction engine developed by KLM in the Google Cloud Platform, with the outputs then streamed straight back to the DMP. From here, the DMP can activate rule-based segments based on the prediction outcome. And by syncing this with DoubleClick, ads can be served and targeted to maximise relevance." [Google, 2017c]

The first documented occurrence of probabilistic client identification is in our view Causata [2010].[111] Causata specialized in storing client identification metadata together with the collected data from end-users. Over the years cross-device tracking technology has developed (see, e. g., Funkhouser, Malloy, Alp, Poon, and Barford [2018], Malloy, Barford, Alp, Koller, and Jewell [2017]). More recent, it has caught the attention of marketers. We recommend to track and replay an interview with Paul Phillips, the founder of Causata, describing the technological challenges of identity association across devices and marketing channels during the development. In particular, we recommend to pay attention to the approach Causata had chosen in the beginning [Greco & Shumpert, 2011]. As a sequel, we refer to Brookman [2015]; Brookman, Rouge, Alva, and Yeung [2017] who presented the

110 In addition to our definition we remark that the recent California Consumer Privacy Act of 2018 [Secretary of State, 2018, AB-375] contains a definition for a probabilistic identifier: „Probabilistic identifier means the identification of a consumer or a device to a degree of certainty of more probable than not based on any categories of personal information included in, or similar to, the categories enumerated in the definition of personal information."

111 The company was acquired by NICE Systems in 2013 (see Table C.2).

topic at the FTC Workshop on cross-device tracking and to Laszlo and Smith [2015, pp. 10–11].[112]

2.4 WEB-BASED FINGERPRINTING

In this section we focus on web-based fingerprinting because it gives a good insight into modern tracking mechanisms and the new detection methods. Knowledge of these mechanisms is important to understand the law and to be able to formulate new rulings. Below we provide two definitions from the literature, one on fingerprint, and one on the process of fingerprinting.

DEFINITION 2.5: „A FINGERPRINT is defined to be a set of information elements that identifies a device or application instance." [Cooper, Tschofenig, Aboba, Peterson, Morris, Hansen, & Smith, 2013, RFC 6973]

DEFINITION 2.6: „FINGERPRINTING is the process of an observer or attacker uniquely identifying (with a sufficiently high probability) a device or application instance based on multiple information elements communicated to the observer or attacker." [Cooper et al., 2013, RFC 6973]

We briefly discuss four types of fingerprinting (Subsection 2.4.1). In Subsection 2.4.2 we briefly address the risks of fingerprinting associated with new technological standards. In Subsection 2.4.3 we show twenty types of common metadata. In Subsection 2.4.4 we provide a reflection on the Anonymous Subscriber Identifier (ASID). Finally, we discuss an example of active fingerprinting: Hotjar (Subsection 2.4.5) and fingerprinting mitigations (Subsection 2.4.6).

2.4.1 *Four types of fingerprinting*

The topic of web-based fingerprinting is actively studied by WPM scholars.[113] We recall Boda, Földes, Gulyás, and Imre [2011] as an

112 Laszlo and Smith [2015, p. 11] noted that „probabilistic identifiers are created by ingesting hundreds of data points and using statistical models to draw conclusions, resulting in a high confidence level for identifying end-users or devices."

113 We refer to, e. g., Al-Fannah, Li, and Mitchell [2018], Gómez-Boix, Laperdrix, and Baudry [2018], Kobusińska, Pawluczuk, and Brzeziński [2018], Vastel, Rudametkin, and Rouvoy [2018], Aleem, Ishtiaq, Abbasi, and Islam [2017], Haga,

early attempt to measure fingerprinting as a substitute for tracking with a persistent UID HTTP cookie. Doty [2017] considers three types of fingerprinting (A, B, and C). Recent research has led to a fourth type of fingerprinting (D). We first list them below and briefly describe them.

(A) passive fingerprinting,
(B) active fingerprinting,
(C) cookie-like fingerprinting, and
(D) cross-device fingerprinting.

A AND B: PASSIVE AND ACTIVE FINGERPRINTING

We remark that passive and active fingerprinting have as their main goal the identification of the end-user without relying on storing or caching persistent UIDs in the browser. This technology is therefore known as a type of *stateless web tracking* (see our technical Definition 2.7 below). An example of active fingerprinting is given in Subsection 2.4.5.

DEFINITION 2.7: STATELESS WEB TRACKING is a type of web tracking that does not rely on storing or caching a persistent UID in the browser.[114]

C: COOKIE-LIKE FINGERPRINTING

In contrast, cookie-like fingerprinting is known as a type of *stateful web tracking* due to its reliance on storing information in the browser (see our technical Definition 2.8 below).[115] Its main goal is the identification of an end-user by a site that first sets and later retrieves a state stored by a user agent or device.

DEFINITION 2.8: STATEFUL WEB TRACKING is a type of web tracking that relies on storing or caching a persistent UID in the browser.[116]

Takata, Akiyama, and Mori [2017], or Nikiforakis, Kapravelos, Joosen, Kruegel, Piessens, and Vigna [2013].

114 Viz. Definition 1.8.

115 From a technical point of view, a web-based fingerprint is not the same as a HTTP cookie (see, e. g., Barth [2011, RFC 6265]).

116 Ibid.

D: Cross-device fingerprinting

Yuan, Maple, Chen, and Watson [2018] reported on a novel technique to track end-users across devices through their (unique) typing behavior. In advance to the paper by Yuan et al. [2018], notable progress in web-based fingerprinting was reported by Cao, Li, and Wijmans [2017] and Van Goethem and Joosen [2017]. They identified a novel browser fingerprint that can identify not only end-users behind one browser, but also end-users that use different browsers on the same machine.[117]

Summary of development: fingerprinting capabilities

Based on the above, we explicitly remark that new types of web-based fingerprinting enable tracking of end-users not only

(1) across different websites,
(2) across different devices (desktop and mobile), but also
(3) across browser sessions on different browser instances on the same device.

2.4.2 *Fingerprinting risks*

Fingerprinting risks are associated with the new browser and device APIs. Olejnik, Acar, Castelluccia, and Diaz [2015] reported quite extensively on fingerprinting associated with the high precision readouts with a capacity of the level of the battery by the Hypertext Text Markup Language (HTML)5 Battery Status API [Kostiainen & Lamouri, 2014]. More recently, W3C published six new sensor APIs as Candidate Recommendations; all containing an extensive section on privacy-consideration.[118]

We remark that besides *web-based* fingerprinting, researchers studied new fingerprinting risks, i. e., *sensor* fingerprinting, due to anomalies in sensors on mobile devices (e. g., due to differences in energy consumption). The anomalies can be used to derive a

117 Cao et al. [2017]: „Our approach adopts OS and hardware levels features including graphic cards exposed by WebGL, audio stack by Audio-Context, and CPU by hardwareConcurrency."

118 See (1) Generic Sensor API [Waldron, Pozdnyakov, & Shalamov, 2018], (2) Ambient Light Sensor [Kostiainen, 2018], (3) Accelerometer [Kostiainen & Shalamov, 2018], (4) Gyroscope [Kostiainen & Pozdnyakov, 2018], (5) Magnetometer [Kostiainen & Bhaumik, 2018], and (6) Orientation Sensor [Pozdnyakov, Shalamov, Christiansen, & Kostiainen, 2018].

(unique) device fingerprint. For instance, fingerprints based on the microphone, the speaker, the motion sensor, the accelerometer, a combination of the gyroscope with inaudible sound, or a combination of the accelerometer and the gyroscope.[119]

2.4.3 *Twenty types of common metadata*

To demonstrate the usefulness of our definitions (Definition 2.3 and Definition 2.4) alongside the definitions for fingerprint (Definition 2.5) and fingerprinting (Definition 2.6), I composed a table of metadata (Table 2.2, next page).[120] The table consists of five categories (Ad, App, Context, Device, End-user) with 20 types of common metadata (see column 1) for RTB client identification. Our table (partly) overlaps with the table by Laszlo and Smith [2015], but is more extensive and compares the application of UIDs between desktop and mobile.

We note that the table is not meant to be complete. Our aim is to deepen the reader's view and understanding. The table provides an overview of currently used metadata for RTB and as such it gives an insight into the current state of tracking technology. For instance, Apple replaced the Apple Unique User Identifier (UUID) with the Identifier For Advertisers (IDFA) with the release of Apple's mobile Operating System (OS) IOS6. Since mobile phones running (older) OS versions still contain the Apple UUID, the identifier was included in the table. Furthermore, we remark that Google replaced the Android ID in 2013 by the Google Anonymous Advertising Identifier (AAID) [Google, n.d.-b].

2.4.4 *Anonymous Subscriber Identifier*

We close our overview by a reflection on the ASID [Rutgers, 2008]. Technically it is a UID injected in a HTTP header field by the end-user's internet service provider. HTTP header injection is thought to be a relatively new phenomenon that caused privacy concerns

119 See, e. g., Das, Acar, Borisov, and Pradeep [2018], Das, Borisov, and Chou [2018], Das, Borisov, and Caesar [2016], Kurtz, Gascon, Becker, Rieck, and Freiling [2016], Bojinov, Michalevsky, Nakibly, and Boneh [2014], Dey, Roy, Xu, Choudhury, and Nelakuditi [2014], or Zhou, Diao, Liu, and Zhang [2014].

120 We remark that a recent study by Gómez-Boix et al. [2018] compared 17 metadata attributes based on Shannon's entropy (see Definition 3.12). See also the early results by Laperdrix, Rudametkin, and Baudry [2016].

Table 2.2: Common metadata for RTB client identification on desktop and mobile.

METADATA	EXAMPLE(S)	DESKTOP	MOBILE
	(1) AD		
Advertising ID	Apple UUID and IDFA	✓	✓
	Android ID and Google AAID		✓
	Microsoft advertising ID	✓	✓
Hash	MD5 or SHA1 hash of the Advertising ID	✓	✓
	WebGL fingerprint, canvas fingerprint	✓	✓
Size	Standard size of the ad	✓	✓
	(2) APP		
Browser	Vendor, version, language, DNT setting	✓	✓
	Browser extensions, user agent string	✓	✓
App Identifier	App ID		✓
	Microsoft Universal App ID	✓	✓
JavaScript	JavaScript support	✓	✓
RESTful postbacks	JSON objects, HTTP cookies	✓	✓
	Mobile-app installation, usage		✓
	(3) CONTEXT		
Bluetooth	Bluetooth MAC Address		✓
Carrier	GSM Cell ID (CID)		✓
IP address	Geolocation lookup	✓	✓
	End-user identification	✓	✓
Lat/long	Lat/long coordinates	✓	✓
Time	Time of day, time zone	✓	✓
Weather	Atmospheric sensor		✓
WiFi	WLAN MAC Address		✓
	(4) DEVICE		
Carrier UID	IMSI, IMEI, ASID		✓
Device	Manufacturer, model		✓
	App-independent ID		✓
	OS, OS version, system language	✓	✓
Screen	screen size, pixel density, color depth	✓	✓
Network	MAC address, IP address	✓	✓
	(5) END-USER		
Credit card	Credit card number	✓	✓
Login	End-user ID, email, mobile number (MSIN)	✓	✓

(see, e. g., Hoffman-Andrews [2014], Mayer [2015b], or Bujlow et al. [2015]). However, the technology has been implemented in the Netherlands since 2008 [Rutgers, 2008].

2.4.5 *Example of active fingerprinting: Hotjar*

An example of an active (web-based) fingerprint is given in Listing 2.1. The name is Hotjar, henceforth we call it: hotjar-script.[121]

Listing 2.1: Hotjar active fingerprint (version 0.7.1).

```
hj.fingerprinter.prototype = { (...)
        a = this.userAgentKey(a),
        a = this.languageKey(a),
        a = this.colorDepthKey(a),
        a = this.timezoneOffsetKey(a),
        a = this.sessionStorageKey(a),
        a = this.localStorageKey(a),
        a = this.indexedDbKey(a),
        a = this.addBehaviorKey(a),
        a = this.openDatabaseKey(a),
        a = this.cpuClassKey(a),
        a = this.platformKey(a),
        a = this.doNotTrackKey(a),
        a = this.pluginsKey(a),
        a = this.adBlockKey(a),
        a = this.hasLiedLanguagesKey(a),
        a = this.hasLiedResolutionKey(a),
        a = this.hasLiedOsKey(a),
        a = this.hasLiedBrowserKey(a),
        l = this.x64hash128(a.join("~~~"), 31));
        return l
},      (...) };
hj.fingerprinter.VERSION = "0.7.1";
return hj.fingerprinter
```

I retained 30 different hotjar scripts spanning from January 2016 to March 2018. Although these scripts have different unique

121 We retained the script 'modules-6081698dc2a04df4b0848520a08b4ffb.js': URL: https://web.archive.org/web/20180331152129/http://script.hotjar.com/modules-6081698dc2a04df4b0848520a08b4ffb.js (31 march 2018).

names, they all contain version 0.7.1 of the same hotjar-script.[122] According to Some, Rezk, and Bielova [2018],[123] the script was found on 10.94% of websites in the top 10,000 Alexa sites in March 2018. Below we will breakdown the script such that the meaning of the JavaScript code is clear.

Unique hash over 18 types of metadata

Surprisingly, the fact that the code is aimed at deriving a fingerprint (Listing 2.1, r. 1) was not obscured, apart from the fact that the code was minified. Minification is known as a process of removing all unnecessary characters from source code without changing its functionality.

The key line of code is Row 20. It contains a reference to a hash function 'x64hash128' which is included in the hotjar-script.[124] In fact, the hash function calculates a UID by a specialized JavaScript hash function over the 18 types of metadata in the lines above (i. e., Listing 2.1, rr. 2–19). For instance, the user agent string (Listing 2.1, r. 2),[125] the language of the browser (Listing 2.1, r. 3),[126] and the color depth of the screen (Listing 2.1, r. 5).[127]

Since it would take too long to discuss each type of metadata one by one, we provide briefly two incisive privacy examples: (A) Do Not Track and (B) ad blocker. We analyze both extractions of metadata below.

A: Do Not Track

The metadata includes the fact whether an end-user has expressed his preference wanting to be tracked or not. The browser's Do

122 The oldest file retained by the Internet Archive is from January 2016 URL: `https://web.archive.org/web/20160128140816/https://script.hotjar.com/modules-a29345a78164b8999baecd49172917fd.js` (31 March 2018).

123 Some et al. [2018] continuously crawl the top 10,000 Alexa sites and provide statistics of, e. g., JavaScript Libraries.

124 The hash function returns a string with a UID, i. e., '("ooooooo" + (g[o] »> o).toString(16)).slice(-8) + ("ooooooo" + (g[1] »> o).toString(16)).slice(-8) + ("ooooooo" + (c[o] »> o).toString(16)).slice(-8) + ("ooooooo" + (c[1] »> o).toString(16)).slice(-8)'.

125 Example user agent string of the browser: 'Mozilla/5.0 (Windows NT 10.0; Win64; x64; rv:59.0) Gecko/20100101 Firefox/59.0'

126 Example language of the browser: 'nl'.

127 Example color depth: '24' bits per pixel.

Not Track (DNT) [Fielding & Singer, 2018] setting is to be seen in Listing 2.1, r. 13.[128]

B: AD BLOCKER

The metadata includes the fact whether the end-user has activated ad blocking (Listing 2.1, r. 15). The function to detect ad blocking is shown in Listing 2.2. What the script does is (1) inserting an element labeled 'ads' in the browser Document Object Model (DOM) (Listing 2.2, rr. 2–3) and (2) simply testing whether the newly created element is blocked or not (Listing 2.2, rr. 4–7).

Listing 2.2: Hotjar: ad-blocking detection.

```
1  getAdBlock: function () {
2          var a = document.createElement("div");
3          a.setAttribute("id", "ads");
4          try {
5                  return document.body.appendChild(a),
6                  document.getElementById("ads") ? !1 : !0
7          } catch (d) {
8                  return !1
9          }
10 }
```

2.4.6 *Fingerprinting mitigations*

Other WPM research is dedicated to mitigate the risk of fingerprinting. For instance, Olejnik et al. [2015], proposed to mitigate the risk by reducing the precision of the readings. Furthermore, Laperdrix, Rudametkin, and Baudry [2015] aimed to modify automatically the fingerprint that a platform exhibits. They developed a prototype called Blink „to experiment the effectiveness of their approach at randomizing fingerprints".[129] Vastel et al. [2018] developed FP-TESTER, a tool to help, amongst others, developers of browser extension to identify fingerprinting risks. See also, e. g., Starov and Nikiforakis [2017b] and Laperdrix [2017].

128 The function was declared as follows. 'getDoNotTrack: function () { return navigator.doNotTrack ? "doNotTrack: " + navigator.doNotTrack : "doNotTrack: unknown"}'

129 URL: `https://github.com/DIVERSIFY-project/blink` (13 November 2017).

Doty [2017] is meant to mitigate browser fingerprinting in web specifications. The W3C draft (a) identifies three different risks on privacy due and (b) formulates eight different best practices to mitigate the privacy impact. Doty [2015] recommended eight technical principles based on data minimization.[130] However, in the European Union a tracking fingerprint has a legal implication. From a policy point of view, a *tracking* fingerprint is a specialized cookie that meets the same legal requirements as a *tracking* cookie (see, e. g., Article 29 Working Party [2014c, WP 224]).

Modern browsers tend to mitigate attempts to fingerprint end-users through misuse of HTML5 API features. For instance, Firefox detects and blocks fingerprinting attempts using the HTML5 canvas element. A second example is the mitigation by Firefox of the risk of fingerprinting based on the types of fonts installed locally in the browser. Advertisers could query the list of fonts and create a UID (fingerprint) for each browser (see, e. g., Kobusińska et al. [2018]).[131]

2.5 HTTP HEADER FIELDS

Originally, HTTP was designed as a communication protocol for hypertext information systems.[132] Yet, it is nowadays also a voluntary global internet standard [Fielding & Reschke, 2014b, RFC 7231].[133] We discuss both issues: (a) the communication protocol in Subsection 2.5.1 and (b) the global internet standard in Sub-

130 The eight principles are: „(1) avoid unnecessary increases to the surface for passive fingerprinting, (2) prefer functionally-comparable designs that don't increase the surface for active fingerprinting, (3) mark features that contribute to fingerprintability, (4) specify orderings and non-functional differences, (5) design apis to access only the entropy necessary, (6) enable graceful degradation for privacy-conscious end-users or implementers, (7) avoid unnecessary new cookie-like local state mechanisms, and (8) highlight any local state mechanisms to enable simultaneous clearing."
[Doty, 2015].

131 Kobusińska et al. [2018, p. 5] address retrieving fonts (1) with the canvas browser object or (2) by querying Cascading Style Sheet (CSS) fonts, to detect differences between texts.

132 HTTP is defined by the HTTP/1.1 syntax [Fielding & Reschke, 2014a, RFC 7230]. Syntactically, HTTP is a formal language [Crocker & Overell, 2008, RFC 5243]. The semantics and content of the HTTP protocol (HTTP/1.1) is defined by Fielding and Reschke [2014b, RFC 7231].

133 For completeness, we remark that Fielding and Reschke [2014a, RFC 7230] contains the message syntax and routing, and that Fielding, Nottingham, and Reschke [2014, RFC 7234] sees to caching of the HTTP protocol (HTTP/1.1).

section 2.5.2. The discussion aims to give the reader some insight into the intricacies that play a key role in privacy discussions.

2.5.1 *HTTP as a protocol*

The main part of the RFC 7231 is about the 'data flow'.[134] Instances are a click on a hyperlink and a movement of the mouse hovering over a picture. All events generate measurements of data which may be sent to a web server. Event tracking relies on an end-user interacting with web content. The flow of event data is communicated with the HTTP protocol.

On a technical level, HTTP is (1) bidirectional, (2) stateless, and (3) operates at the application level. Semantically, HTTP is summarized as follows.

> „This document [RFC 7231] defines the semantics of HTTP/1.1 messages, as expressed by request methods, *request header fields*, response status codes, and *response header fields*, along with the payload of messages (metadata and body content) and mechanisms for content negotiation." [Fielding & Reschke, 2014b, RFC 7231, Abstract] (emphasis added)

We remark that HTTP header fields are key elements of the protocol. In the following we will first discuss the difference between (A) the HTTP request header and (B) the HTTP response header. Next, we briefly discuss (C) header field registration. We end this subsection with (D) a subsection conclusion.

A: HTTP REQUEST HEADER

The Internet Assigned Numbers Authority (IANA) maintains a list of registered HTTP header fields [IANA, 2015a; 2015b]. The HTTP request header field and the HTTP response header field contain control data and resource metadata. The difference between the two HTTP header fields is as follows.

The HTTP request header [Fielding & Reschke, 2014b, RFC 7231, Section 5] has five different functions,[135]

134 See, e. g., Tschofenig [2013] or Rescorla and Internet Architecture Board [2005, RFC 4101]

135 Gastineau [2015] gives a technical explanation for the functions: „A *POST* request writes to a web server, a *GET* request reads from the server, a *HEAD* request

(1) to provide information about the request context,
(2) to make the HTTP request conditional based on the target resource state,
(3) to suggest preferred formats for the HTTP response,
(4) to supply authentication credentials, or
(5) to modify the processing of the expected HTTP request.

B: HTTP RESPONSE HEADER

The HTTP response header [Fielding & Reschke, 2014b, RFC 7231, Section 7] has two different functions. It provides information about (1) the server and (2) further access to the target resource, or related resources. Although the two types of header fields above differ in the types of information which they provide, we remark that both can be used to communicate (unique) identifiers.

C: HEADER FIELD REGISTRATION

HTTP header fields are registered in the Message Headers registry [Klyne, Nottingham, & Mogul, 2004, RFC 3864]. However, there is no enforcement or legal ruling for registration of newly defined HTTP headers. Moreover, it is not possible to syntactically distinguish between standard and non-standard HTTP header fields. Each has its own level of standardization. Any HTTP header not registered in the IANA list [IANA, 2015a] simply reflects the degree to which people agree on its intended implementation.

D: SUBSECTION CONCLUSION

The main conclusion of this subsection is that, it makes (more) sense to distinguish between registered and unregistered HTTP header fields than between standard and non-standard header fields.[136]

retrieves only the HTTP request header, and a *PUT* request creates and replaces resources on a web server."

136 I am indebted to Fielding [2015], D. Singer [2015], and Tschofenig [2015] who confirmed this insight independently from one another. They were, of course, considering this as experts, in their personal capacity.

2.5.2 *HTTP as an internet standard*

HTTP is a core part of today's internet and it plays an important role in the internet governance model.[137] The playing field consists of a number of key organizations. We discuss the role of seven of these organizations in relation to the HTTP header fields (denoted by number*).

(1*) Internet Assigned Numbers Authority (IANA)
(2*) Internet Engineering Task Force (IETF)
(3*) Internet Corporation for Assigned Names and Numbers (ICANN)
(4*) Internet Society (ISOC)
(5*) Internet Architecture Board (IABoard)
(6*) Internet Engineering Steering Group (IESG)
(7*) World Wide Web Consortium (W3C)
(8*) National Telecommunications and Information Administration (NTIA)

To make sure that the web keeps running, the HTTP standard is coordinated by the IANA (1*) in conjunction with the IETF (2*). ICANN (3*) is responsible for performing the IANA functions, e. g., maintaining key Internet protocols, allocating and administering IP addresses, port numbers, and the time-zone database.[138] The IETF is responsible for the development of HTTP under the auspices of the ISOC (4*). Within the IETF, the Internet Architecture Board (5*) and the IESG (6*) coordinate the creation of internet standards and the RFC document series.

Other organizations such as the W3C (7*) contribute to HTTP functionality. A recent example is the Do Not Track (DNT) request

137 For a brief description of how the internet works, we refer to the W3C wiki: URL: `https://www.w3.org/wiki/How_does_the_Internet_work` (19 November 2017).

138 The IANA functions are as follows: „When you want to visit a website, you type or paste the site's domain name into your browser, or click on an html link. That domain name is sent to a server which translates the name into a series of numbers - the Internet Protocol or IP Address - which the server uses to direct your request to the website's physical location. This all happens in the blink of an eye. Those names and numbers are called 'UIDs' and are aligned with a standard set of protocol parameters that ensure computers can talk to and understand each other. These are part of the IANA functions, managed by ICANN, the Internet Corporation for Assigned Names and Numbers. These functions aren't just limited to browsing the internet - they also enable you to send an email or backup photos to the cloud, amongst other tasks." URL: `https://www.icann.org/en/system/files/files/functions-basics-07apr14-en.pdf` (13 November 2017)

header field, a HTTP header mechanism for expressing the end-user's preference regarding tracking [Doty et al., 2016; Fielding & Singer, 2018]. The W3C and the IETF run side by side because both develop internet standards. However, the web is more at the center of the W3C.

Here, we remark that the internet governance model was four years ago in a state of flux (see, e. g., Clark [2015]). The United States government's NTIA (8*) governed the internet Domain Name System (DNS) through a contract with the ICANN.[139] Then the United States government handed over its governance of the IANA functions to an international multi-stakeholder community. ICANN nowadays performs the IANA functions.[140] For historical reasons we mention that the process was completed on 30 September 2016.[141]

2.6 FOUR EXPERIMENTAL FRAMEWORKS

In this section we focus on four experimental frameworks in the field of WPM. Below, we briefly discuss the characteristics of each of them: Fourthparty (Subsection 2.6.1), FPDetective (Subsection 2.6.2), OpenWPM (Subsection 2.6.3), and WebXray (Subsection 2.6.4).

2.6.1 *Fourthparty*

Mayer and Mitchell [2012] developed FourthParty, a Firefox extension that they used to record the communication to the different tracking parties. The *technical* focus of FourthParty was mainly on explicitly assigned client-side identifiers. We remark that the focus relates to the first category identified by Janc and Zalewski [2014] (Section 2.2). The *policy* focus of the framework was measurement of effectiveness of DNT usage by publishers.

139 The scope of the IANA contract included: „(1) the coordination of the assignment of technical internet protocol parameters; (2) the administration of certain responsibilities associated with internet DNS root zone management; (3) the allocation of internet numbering resources; and (4) other services related to the management of the .ARPA and .INT top-level domains." URL: `https://www.ntia.doc.gov/page/iana-functions-purchase-order` (21 August 2017).

140 See, e. g., URL: `https://www.ntia.doc.gov/blog/2016/update-iana-transition` (21 August 2017) or URL: `https://www.internetsociety.org/iana-transition/` (13 November 2017).

141 For a timeline, see URL: `https://www.internetsociety.org/ianatimeline/` (13 November 2017).

FourthParty used Selenium [Selenium Project, 2004] to automate site visits. FourthParty was the first attempt to standardize WPM-research data in an SQLite [Hipp, 2000] database schema (see Eubank et al. [2013, p. 4]) which allows sharing between researchers. We note that the need for sharing research data in a standardized format was one of the outcomes of the discussions during the workshop web privacy measurement hosted by the Berkeley Center for Law & Technology, in 2012.

Two studies are based on the FourthParty experimental framework: (1) Chaabane, Ding, Dey, Kaafar, and Ross [2014, pp. 4–5] measured leakage of personal data by 377 apps in social networks with a modified version of FourthParty and (2) Eubank et al. [2013] ran six 500-site [Alexa, n.d.] web crawls on different hardware (i. e., PC, smartphone, and tablet).

2.6.2 *FPDetective*

Whereas Mayer and Mitchell [2012] had a focus on persistent browser data (e. g., HTTP cookies and UIDs retained in the browser cache and local storage) Acar, Juarez, Nikiforakis, Diaz, Gürses, Piessens, and Preneel [2013] zoomed in on a particular type of tracking, i. e., canvas fingerprinting (see also, e. g., Raschke and Küpper [2018]). They presented the results of a first large-scale WPM study of 100,000 websites and found that it was present on 5% of the Alexa [n.d.] top 100,000 websites.

Building on the lessons-learned of FourthParty by automating a Firefox browser with Selenium [Selenium Project, 2004], they crawled websites in parallel with up to 30 browsers [Acar et al., 2013, p. 678]. They used a combination of three data collection methods:

(1) a browser modification to collect JavaScript data (e. g., the line number of the calling JavaScript code),
(2) persistent browser data retained in the end-user's profile folder, and
(3) a capture of the network HTTP data with Mitmproxy [Cortesi, Hils, Kriechbaumer, & contributors, 2010].

The logs of these data collection methods were retained in an SQLite [Hipp, 2000] database schema for further analysis.

2.6.3 *OpenWPM*

OpenWPM [Englehardt et al., 2014] builds on FourthParty and FPDetective (cf. [Narayanan & Reisman, 2017, p. 4]). It is the first WPM framework that scales to at least 1 million websites [Englehardt & Narayanan, 2016]. In fact, the framework is sufficiently stable and scalable for continuous crawling. Researchers can modify the OpenWPM framework with specialized source code to fit their needs (see, e. g., Englehardt [2018], Englehardt, Han, and Narayanan [2018], and Englehardt, Reisman, Eubank, Zimmerman, Mayer, Narayanan, and Felten [2015].[142] WPM scholars have modified OpenWPM, e. g., to detect a cookie banner.[143] Although OpenWPM can be modified, there is no official notion of a plugin.[144]

OpenWPM retains the logs in an SQLite database schema. It builds on the FourthParty database schema and the philosophy of sharing research data between researchers for further analysis. For instance, Van Eijk and Toubiana [2016] wrote a script to convert research data, such as the 1 million dataset [Englehardt, 2016], from the OpenWPM format to the Neo4j [Efrem, Svensson, & Neubauer, 2000] graph database for further analysis.[145]

Some researchers, e. g., Maass, Wichmann, Pridöhl, and Herrmann [2017] and Andersdotter and Jensen-Urstad [2016] have proposed automated WPM-scanning services for use by, e. g., the supervisory authorities to perform regularly scheduled compliance checks.

It is noted that under Article 77, GDPR [Parliament of the EU and the Council, 2016, 2016/679] end-users have the right to lodge

142 Englehardt et al. [2018] investigated emails from commercial mailing-lists and found a network of hundreds of third parties that track email recipients via the same tracking techniques used on web pages.

143 See URL: `https://github.com/citp/OpenWPM/pull/159` (5 April 2018).

144 I am indebted to Narayanan [2018]: „Plugins are whatever source code modifications researchers make to tackle their specific research tasks. While we know that many people have made such modifications, we haven't tracked those in a central way."

145 We built on the lessons learned from a similar script by Van Eijk and Terra [2013] converting Mitmproxy data to Neo4j.

a complaint with a supervisory authority.[146] In fact, Recital 141, GDPR explains the empowerment as follows.

> „Every data subject should have the right to lodge a complaint with a single supervisory authority, in particular in the Member State of his or her habitual residence, and the right to an effective judicial remedy in accordance with Article 47 of the Charter if the data subject considers that his or her rights under this Regulation are infringed or where the supervisory authority does not act on a complaint, partially or wholly rejects or dismisses a complaint or does not act where such action is necessary to protect the rights of the data subject. The investigation following a complaint should be carried out, subject to judicial review, to the extent that is appropriate in the specific case. (...)"

The complaint has to be investigated by the supervisory authority within a short time frame (viz. Article 78 sub 2, GDPR).[147]

Moreover, the end-user has the right to mandate a privacy organization to lodge the complaint on his or her behalf (viz. Article 80 GDPR and Recital 142 GDPR).[148]

Based on the above, we may conclude that automated WPM-scanning services clearly empower data subjects and privacy or-

146 Article 77 sub 1, GDPR: „Without prejudice to any other administrative or judicial remedy, every data subject shall have the right to lodge a complaint with a supervisory authority, in particular in the Member State of his or her habitual residence, place of work or place of the alleged infringement if the data subject considers that the processing of personal data relating to him or her infringes this Regulation." Article 77 sub 2, GDPR: „The supervisory authority with which the complaint has been lodged shall inform the complainant on the progress and the outcome of the complaint including the possibility of a judicial remedy pursuant to Article 78."

147 Article 78(2) GDPR: „Without prejudice to any other administrative or non-judicial remedy, each data subject shall have the right to a an effective judicial remedy where the supervisory authority which is competent pursuant to Articles 55 and 56 does not handle a complaint or does not inform the data subject *within three months* on the progress or outcome of the complaint lodged pursuant to Article 77." (emphasis added)

148 Article 80 GDPR: „The data subject shall have the right to mandate a not-for-profit body, organisation or association which has been properly constituted in accordance with the law of a Member State, has statutory objectives which are in the public interest, and is active in the field of the protection of data subjects' rights and freedoms with regard to the protection of their personal data to lodge the complaint on his or her behalf, to exercise the rights referred to in Articles 77, 78 and 79 on his or her behalf, and to exercise the right to receive compensation referred to in Article 82 on his or her behalf where provided for by Member State law."

ganizations, e. g., by scanning for possible violations of the law within a specific sector. For instance, Andersdotter and Jensen-Urstad [2016] addressed compliance by municipalities in Sweden.

2.6.4 *WebXray*

WebXray [Libert, 2015c] is a specialized WPM framework with a focus on the question: who is the owner of a domain? Data is collected by visiting a list of websites provided by the researcher. Data is stored in a database and attributed with information about domain ownership ([Libert, 2015a]).[149] The metadata attribution was compiled manually from various sources, such as WHOIS data [Daigle, 2004, RFC 3912] and developer documentation.[150] Due to the static nature of identification the WebXray framework has its limits in the fast changing landscape due to mergers, acquisitions, and partnerships.

WebXray is fit for the purpose of for more complex automated analysis. For instance, Libert [2018] created an add-on for the framework to automate auditing privacy policies. Finally, we mention the following four studies based on WebXray revealing the monitoring of end-user behavior in different contexts.

(1) Feng, Vasconcellos Vargas, Sakurai, Yu, and Ishikawa [2017] crawled 222 websites for jobs and career opportunities,
(2) Hauschke [2016] analyzed 4,753 websites of libraries in Germany,
(3) Libert [2015a] analyzed the top 1 million websites [Alexa, n.d.] and traced back 500 distinct domains to 140 different companies. At least 100 domains were matched manually by using WHOIS-domain lookups or the website '`crunchbase.com`', and
(4) Libert [2015b] also analyzed 80,124 health-related web pages and found that the full HTTP referrer string poses a privacy risk when it exposes information about specific conditions, treatments, and diseases.[151]

149 Similarly, in an attempt to build a picture of the American digital news publishing ecology Lindskow [2016] compiled an extensive database on business units, owners, purposes, and industries (see also, Sørensen and van den Bulck [2018]).

150 Domain ownership is determined on the parent domain of the Fully Qualified Domain Name (FQDN), e. g., '2mdn.net' which belongs to Google [DoubleClick]. HTTP requests to sub-domains, e. g., '0.2mdn.net', are ignored.

151 The web pages were derived from the 50 top search results for 1,986 common diseases.

2.7 OVERVIEW OF TEN WPM VISUALIZATION TOOLS

Ad technology has been unraveled to a certain extent by transparency tools. We refer to the following ten (well-known) tools.

(1) Lightbeam [Mozilla, 2011].[152]
(2) Web Tracking Detection System (WTDS) [Van Eijk, 2011a].
(3) Cookie-miner [Pannetrat, 2012].[153]
(4) CookieViz [Petitcolas, 2013].
(5) TrackerMap [Evidon, 2013; Nielsen, DeMille, Kilrain, Meyer, Donohoo, Kozek, Van Oss, & Signanini, 2016].
(6) Mobilitics [Baudot, Delcroix, Le Grand, Petitcolas, Jagdish, Castelluccia, Lefruit, & Roca, 2012].
(7) MobileScope [Soltani, Cortesi, & Campbell, 2012].[154]
(8) Netograph [Cortesi, 2011; 2017].
(9) Request Map Generator [Hearne, 2013].
(10) TrackerScope [O'Neill, 2015].

The first five tools can be denoted as early attempts to increase transparency of online advertising. Lightbeam, WTDS, Cookie-miner, CookieViz, and TrackerMap focus primarily on the interconnectedness of (ad tech) companies when a researcher visits different websites. The data is collected with a focus on the *browser*.

In contrast, Mobilitics and MobileScope have a different focus. These two tools focus on data leakage of *mobile apps* instead of the *browser*.[155] Gaining insight into the data leakage has contributed to our understanding of opaque dataflows on smart mobile devices.

The remaining tools, i. e., Netograph, Request Map Generator, and TrackerScope, deserve special attention. We consider them

152 Lightbeam was developed by Mozilla as an extension for its Firefox browser. Lightbeam started as 'Collusion' in 2011 and renamed to 'Lightbeam'.

153 Pannetrat [2016] wrote a tool called Cookie-miner „that is the 'father' or 'grandfather' of CookieViz."

154 MobileScope was developed to developed to test personal-information leakage of mobile apps. The specialized software processes the HTTP headers in real-time and filters, e. g., the Unique Device Identifier (UDID), the mobile device's Media Access Control (MAC) address, the latitude, the longitude, the end-user's email address and other (deterministic) personal identifiers.

155 See Table 2.2. Technically speaking, the browser on a smart mobile device is also an app. However, the browser is a special case as it can store HTTP cookies. Mobile apps cannot store HTTP cookies and have to rely on other mechanisms to recognize an end-user over time.

breakthroughs in WPM. All three tools aim to increase transparency in order to deepen our understanding of how ad tech works. In the following we will briefly discuss them.

Netograph, Request Map Generator, and TrackerScope are special cases. They are implemented as a web service that crawls a webpage. These tools split a *webpage* up into the page's various HTTP elements and indicate how these elements are related to web tracking.[156] Netograph has a focus on different (persistent) storages, e. g., HTTP cookie, flash Local Shared Object (LSO), HTML5 local storage. Request Map Generator is an extension to the performance-testing web service WebPageTest [Hearne, 2008] and has a focus on the time taken to load a webpage.[157] TrackerScope has a focus on the extent to which websites implemented the technical building blocks of the W3C Do Not Track (DNT) protocol.

In summary, we may state that web-tracking graph visualizations contribute to our understanding of how and to what extent event-tracking data is collected. The tools explore (currently still) opaque data flows from ad technology from different perspectives, i. e.,

(1) browser,
(2) mobile apps,
(3) webpage (performance),[158] and
(4) advertisements.

2.8 CHAPTER CONCLUSION

At the end of our overview, we arrive at five conclusions.

(1) HTTP cookies and common device metadata enable tracking of end-users across the web.
(2) The HTTP cookie is still considered as the dominant identifier for *desktop* and *mobile web* advertising.[159]

156 urlQUery.net is a tool with a similar approach. It is a service for detecting and analyzing web-based malware. „It provides detailed information about the activities a browser does while visiting a site and presents the information for further analysis." URL: `https://urlquery.net/index.php` (5 July 2016).

157 See also Hearne [2015] for an informative description of the capabilities of Request Map Generator.

158 In addition to Hearne [2013] (Request Map Generator), we refer to e. g., Yu, Macbeth, Modi, and Pujol [2016] who deeply investigated the impact on the performance of a webpage by third-party resources.

159 I am indebted to Mitchell [2016] who confirmed the insight.

(3) As a consequence, current ad technology for *desktop* advertising still relies on cookie syncing.
(4) In contrast, the device identifier is considered the dominant identifier for *mobile app* advertising on smart mobile phones and tables.
(5) As new HTML5 APIs are created, new fingerprinting risk will emerge.[160]

In this chapter we saw that web-tracking activities can be told as a story. The story resides in the network-flow data. Then we gave a description (it is a glimpse of a description) why current WPM methodologies still do not support our view in full detail. The main question as formulated in our problem statement is whether we can describe the whole story, i. e., measuring the privacy component connected to Real-Time Bidding? Whatever the case, we are now well equipped to address RQ1 and RQ2. In Chapter 3 we define and apply our approach GBMA (Graph-Based Methodological Approach) to examine the path from data collection to graph analysis. Thereafter, in Chapter 4, we will direct our attention to RQ2 to investigate with GBMA the picture that emerges with multiple actors and their interrelationships in the RTB landscape.

160 The W3C Device and Sensors Working Group (DAS WG) creates client-side APIs „that enable the development of web applications that interact with device hardware, sensors, services and applications such as the camera, microphone, proximity sensors, native address books, calendars and native messaging applications." The road map of the working group contains specifications for (1) environment sensors, (2) motion sensors, (3) devices, e. g., the Battery Status API, and (4) the Sensor Framework, i. e., the Generic sensor API. URL: `https://www.w3.org/2009/dap/` (13 November 2017).