



Universiteit
Leiden
The Netherlands

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

Eijk, R.J. van

Citation

Eijk, R. J. van. (2019, January 29). *Web privacy measurement in real-time bidding systems. A graph-based approach to RTB system classification*. Retrieved from <https://hdl.handle.net/1887/68261>

Version: Not Applicable (or Unknown)

License: [Licence agreement concerning inclusion of doctoral thesis in the Institutional Repository of the University of Leiden](#)

Downloaded from: <https://hdl.handle.net/1887/68261>

Note: To cite this publication please use the final published version (if applicable).

Cover Page



Universiteit Leiden



The following handle holds various files of this Leiden University dissertation:

<http://hdl.handle.net/1887/68261>

Author: Eijk, R.J. van

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

Issue Date: 2019-01-29

Web Privacy Measurement in Real-Time Bidding Systems

A Graph-Based Approach to RTB system classification

Web Privacy Measurement in Real-Time Bidding Systems

A Graph-Based Approach to RTB system classification

PROEFSCHRIFT

ter verkrijging van
de graad van Doctor aan de Universiteit Leiden,
op gezag van Rector Magnificus prof. mr. C.J.J.M. Stolker,
volgens besluit van het College voor Promoties
te verdedigen op dinsdag 29 januari 2019
klokke 11.15 uur

door

Robbert Jan-Willem van Eijk

geboren te Rosmalen

in 1969

Promotor: Prof. dr. H.J. van den Herik
Co-promotor: Dr. M. Dechesne

Promotiecommissie: Prof. dr. S. van der Hof
Prof. dr. J.N. Kok (Universiteit Twente /
Universiteit Leiden)
Prof. dr. E.O. Postma (Tilburg University)
Prof. dr. P.P. Swire (Georgia Institute of
Technology, Atlanta, USA)
Dr. B.H.M. Custers



SIKS Dissertation Series No. 2019-01

The research reported in this thesis has been carried out under the auspices of SIKS, the Dutch Research School for Information and Knowledge Systems.

ISBN: 978-94-028-1323-4
NUR: 821

© 2019, R.J.W. van Eijk, Creative Commons BY-NC-SA 4.0

Preface

The primary motive for writing a doctoral thesis arose from the wish to investigate Real-Time Bidding (RTB) at the intersection of technology and policy. I wanted to deepen my understanding of what *actually* happens when an end-user visits websites containing advertisements that seem to follow you. Four milestones laid the foundation for this thesis:

- (1) my master's thesis „Web Tracking Detection System: An effective strategy to reduce systematic monitoring and profiling of user habits across websites" [Van Eijk, 2011b];
- (2) I was interviewed for an article on RTB in the New York Times „You for sale: Your attention, bought in an instant" [N. Singer, 2012a; 2012b; 2012c];
- (3) the article led to an item on RTB on the Dutch television [Kamphuis, 2013a; 2013b], and
- (4) the first official draft of the working paper „Web Tracking and Privacy: Respect for context, transparency and control remains essential" [IWGDPT, 2013; 2014], which led to the joint resolution at the 35th International Conference of Data Protection and Privacy Commissioners [ICDPPC, 2013].

I conducted my research at (a) eLaw - Center for Law and Digital Technologies of the Institute for the Interdisciplinary Study of the Law, Faculty of Law of Leiden University and (b) the Dual PhD Centre, Faculty of Governance and Global Affairs of Leiden University. What binds the doctoral candidates at the Dual PhD Centre is that we combine a PhD with a job.

I hope you will find the thesis clear and interesting. I wish you a pleasant reading experience.

The Hague, November 2018

Robbert J. van Eijk

Contents

PREFACE	5
CONTENTS	7
LIST OF ABBREVIATIONS	11
LIST OF DEFINITIONS	14
LIST OF FIGURES	16
LIST OF TABLES	17
LIST OF LISTINGS	18
LIST OF ALGORITHMS	18
LIST OF EQUATIONS	18
1 DATA-DRIVEN ADS	19
1.1 Real-time bidding	20
1.2 Privacy and web tracking	21
1.3 Data	30
1.3.1 Big data	31
1.3.2 Small data	36
1.3.3 Metadata	38
1.4 Problem statement and research questions	41
1.5 Research methodology	42
1.6 Structure of the thesis	43
2 LITERATURE REVIEW	45
2.1 Web privacy measurement	45
2.1.1 Why web privacy measurement matters	47
2.1.2 Event tracking	49
2.1.3 Example of an action: Google Analytics	52
2.1.4 Information asymmetry	53
2.1.5 Multidisciplinary legal-science paradigm	53
2.2 Three client identification mechanisms	56
2.2.1 Technical analysis	56
2.2.2 Five categories of tracking mechanisms	56
2.3 Deterministic and probabilistic identification	57
2.3.1 Deterministic identification	58
2.3.2 Probabilistic identification	58

2.4	Web-based fingerprinting	60
2.4.1	Four types of fingerprinting	60
2.4.2	Fingerprinting risks	62
2.4.3	Twenty types of common metadata	63
2.4.4	Anonymous Subscriber Identifier	63
2.4.5	Example of active fingerprinting: Hotjar	65
2.4.6	Fingerprinting mitigations	67
2.5	HTTP header fields	68
2.5.1	HTTP as a protocol	69
2.5.2	HTTP as an internet standard	71
2.6	Four experimental frameworks	72
2.6.1	Fourthparty	72
2.6.2	FPDetective	73
2.6.3	OpenWPM	74
2.6.4	WebXray	76
2.7	Overview of ten WPM visualization tools	77
2.8	Chapter conclusion	78
3	FROM DATA COLLECTION TO GRAPH ANALYSIS	81
3.1	Data collection	82
3.1.1	Browser data collection	82
3.1.2	Data-flow collection: existing methods	87
3.1.3	Our 3-step procedure of data-flow collection	89
3.1.4	Proxy collection	97
3.2	Data reduction	102
3.3	Data modeling	105
3.3.1	Generic web-tracking model	105
3.3.2	Formalization of graph-mining rules	106
3.3.3	Extended web-tracking model	111
3.3.4	High-entropy identifiers	113
3.3.5	A small-data model	118
3.4	Chapter conclusions	129
3.5	An answer to research question 1	129
4	CATEGORIZATION OF RTB SYSTEMS	131
4.1	The network of partners	133
4.1.1	Acquisitions	134
4.1.2	Mergers	135
4.1.3	Strategic partnerships	135
4.1.4	The Rubicon Project as an example	137
4.1.5	Section conclusions	139
4.2	A definition for RTB system	140

4.3	Theoretical view of RTB systems	140
4.3.1	Inventory sources	142
4.3.2	Publisher	143
4.3.3	Supply Side Platform	143
4.3.4	Demand Side Platform	144
4.3.5	Data Management Platform	145
4.3.6	RTB protocol	146
4.3.7	Private deal	148
4.3.8	Tag-Based Integration	149
4.3.9	Nine steps in the RTB-bidding process	150
4.3.10	Section conclusions	152
4.4	Empirical view of RTB systems	153
4.4.1	Digital media as a contextual data source	153
4.4.2	Stateless deep crawl: EU Feeds	154
4.4.3	Construction of the graph	156
4.4.4	Three empirical observations	160
4.4.5	Section conclusions	170
4.5	Consent as a privacy component	171
4.5.1	Consent as a legal basis under the GDPR	171
4.5.2	Consent under the EPD/EPR	172
4.5.3	Two types of consent implementations	174
4.5.4	Analysis of consent implementations	175
4.5.5	Section conclusions	179
4.6	Graph analysis of partner networks	180
4.6.1	Cluster-edge betweenness	181
4.6.2	Node betweenness	184
4.6.3	Categorization into four types of nodes	185
4.6.4	Section conclusions	191
4.7	Chapter conclusions	192
4.8	An answer to research question 2	195
5	CONCLUSIONS	199
5.1	Answers to the research questions	199
5.2	Answer to the problem statement	200
5.3	Fundamental data protection rights	202
5.4	Further research	205
	REFERENCES	207
	APPENDICES	251
	A: WORKSHOP PARTICIPATION	252
	A1: PARTICIPATING ROLES	252

A2: WORKSHOP OVERVIEW	255
B: FILTERING CHALLENGING PROBLEMS	258
C: ACQUISITIONS, MERGERS, AND PARTNERSHIPS	259
D: PARTNER NETWORKS: DENMARK (DK)	266
SUMMARY	275
SAMENVATTING (SUMMARY IN DUTCH)	279
ACKNOWLEDGMENTS	283
PUBLICATIONS AND PRESENTATIONS	285
CURRICULUM VITAE	289
INDEX	291

List of abbreviations

AAID	- Anonymous Advertising Identifier
ADSL	- Asymmetric Digital Subscriber Line
AGM	- Annotated Graph Mining
AIC	- Akaike Information Criterion
API	- Application Programming Interface
Art. 29 WP	- Article 29 Data Protection Working Party
ASID	- Anonymous Subscriber Identifier
AWS	- Amazon Web Services
CDN	- Content Delivery Network
CFO	- Call For Objections
CFREU	- Charter of Fundamental Rights of the European Union
CID	- GSM Cell ID
COM	- .COM Top Level Domain
CPA	- Cost Per Acquisition
CPC	- Cost-Per-Click
CPM	- Cost-Per-Mille
CSS	- Cascading Style Sheet
CTPC	- Counting Third-Party Cookies
CTR	- Click Through Rate
DMP	- Data Management Platform
DNS	- Domain Name System
DNT	- Do Not Track
DOM	- Document Object Model
DPA	- Data Protection Authority
DSP	- Demand Side Platform
ECHR	- European Convention on Human Rights
ECTHR	- European Court of Human Rights
EDPS	- European Data Protection Supervisor
EFF	- Electronic Frontier Foundation
eLaw	- Center for Law and Digital Technologies
EPD	- e-Privacy Directive
EPR	- e-Privacy Regulation
EU	- European Union
FISA	- Foreign Intelligence Surveillance Act
FQDN	- Fully Qualified Domain Name
FTC	- Federal Trade Commission

GBMA	- Graph-Based Methodological Approach
GBMA	- GraafgeBaseerde Methodologische Aanpak
GDPD	- General Data Protection Directive
GDPR	- General Data Protection Regulation
GLM	- Generalized Linear Model
GPEN	- Global Privacy Enforcement Network
HAR	- HTTP Archive
HSTS	- HTTP Strict Transport Security
HTML	- Hypertext Text Markup Language
HTTP	- Hypertext Transfer Protocol
HTTPS	- HTTP Secure
IABoard	- Internet Architecture Board
IABureau	- Interactive Advertising Bureau
IANA	- Internet Assigned Numbers Authority
IAPP	- International Association of Privacy Professionals
ICANN	- Internet Corporation for Assigned Names and Numbers
ICDPPC	- International Conference of Data Protection and Privacy Commissioners
ICT	- Information and Communication Technology
ID	- Identifier
IDFA	- Identifier For Advertisers
IESG	- Internet Engineering Steering Group
IETF	- Internet Engineering Task Force
IGF	- Internet Governance Forum
IMEI	- International Mobile Equipment Identifier
IMSI	- International Mobile Subscriber Identifier
IP	- Internet Protocol
IPEN	- Internet Privacy Engineering Network
ISOC	- Internet Society
IT	- Information Technology
ITP	- Intelligent Tracking Prevention
IViR	- Institute for Information Law
IWGDPT	- International Working Group on Data Protection in Telecommunications
JICWEBS	- Joint Industry Committee for Web Standards
JNLP	- Java Network Launching Protocol
JSON	- JavaScript Object Notation
LCDS	- Leiden Centre of Data Science
LIACS	- Leiden Institute of Advanced Computer Science
LSO	- Local Shared Object
MAC	- Media Access Control

MEP	- Member of the European Parliament
MIT	- Massachusetts Institute of Technology
MSIN	- Mobile Subscription Identification Number
N/A	- Not Available
NL	- .NL Top Level Domain
NSA	- National Security Agency
NTIA	- National Telecommunications and Information Administration
OBA	- Online Behavioral Advertising
OECD	- Organisation for Economic Co-Operation and Development
ORCID	- Open Researcher and Contributor Identifier
OS	- Operating System
PHP	- Hypertext Preprocessor (earlier: Personal Home Page)
PS	- Problem Statement
RESTful	- Representational State Transfer
RFC	- Request For Comments
RQ	- Research Question
RSS	- Rich Site Summary
RTB	- Real-Time Bidding
SIKS	- Dutch Research School for Information and Knowledge Systems
SOP	- Same Origin Policy
SPDY	- Speedy (networking protocol)
SSP	- Supply Side Platform
TBI	- Tag-Based Integration
TFEU	- Treaty on the Functioning of the European Union
TLS	- Transport Layer Security
UDID	- Unique Device Identifier
UID	- Unique Identifier
URI	- Uniform Resource Identifier
URL	- Universal Resource Locator
UUID	- Unique User Identifier
W3C	- World Wide Web Consortium
WLAN	- Wireless Local Area Network
WPM	- Web Privacy Measurement
WTDS	- Web Tracking Detection System

List of definitions

Definition 1.1	Real-time bidding	20
Definition 1.2	Real-time bidding [Forrester, n.d.]	20
Definition 1.3	Real-time bidding [Google, n.d.]	21
Definition 1.4	Real-time bidding [IABureau, 2012]	21
Definition 1.5	Privacy [Westin, 1967]	22
Definition 1.6	Web tracking [Van Eijk, 2011b]	25
Definition 1.7	Tracking [Fielding & Singer, 2018]	27
Definition 1.8	Web tracking	28
Definition 1.9	Profiling [Parliament of the EU and the Council, 2015]	29
Definition 1.10	Data [Meesters, 2014]	30
Definition 1.11	Big data [Lynch, 2008]	31
Definition 1.12	Big data [Jacobs, 2009]	32
Definition 1.13	Big data [Loukides, 2010]	32
Definition 1.14	Big data [Beyer & Douglas, 2012]	32
Definition 1.15	Big data [T. White, 2012]	33
Definition 1.16	Big data [Mayer-Schönberger & Cukier, 2013]	33
Definition 1.17	Big data [P. Phillips, 2015]	33
Definition 1.18	Small data	36
Definition 1.19	Cookie [Kristol & Montulli, 1997, RFC 2109]	37
Definition 1.20	Metadata	40
Definition 1.21	Identifier [ISO, 2011, 24760-1]	40
Definition 1.22	Identity [ISO, 2011, 24760-1]	41
Definition 2.1	Web privacy measurement [Englehardt et al., 2014]	45
Definition 2.2	Event tracking	49
Definition 2.3	Deterministic client identification	58
Definition 2.4	Probabilistic client identification	58
Definition 2.5	Fingerprint [Cooper et al., 2013, RFC 6973]	60
Definition 2.6	Fingerprinting [Cooper et al., 2013, RFC 6973]	60
Definition 2.7	Stateless web tracking	61
Definition 2.8	Stateful web tracking	61

Definition 3.1	Shallow crawl	92
Definition 3.2	Deep crawl	92
Definition 3.3	Graph [Riesen et al., 2010]	106
Definition 3.4	Web tracking	106
Definition 3.5	Subgraph [Riesen et al., 2010]	108
Definition 3.6	Subgraph A	109
Definition 3.7	Subgraph B	109
Definition 3.8	Subgraph C	109
Definition 3.9	Subgraph D	109
Definition 3.10	Subgraph E	110
Definition 3.11	Induced subgraph [Riesen et al., 2010]	110
Definition 3.12	Shannon entropy [Shannon, 1948]	114
Definition 3.13	High-entropy identifier	114
Definition 3.14	False positive	116
Definition 3.15	False negative	117
Definition 3.16	Graph isomorphism [Riesen et al., 2010]	119
Definition 3.17	Subgraph isomorphism [Riesen et al., 2010]	119
Definition 3.18	HTTP redirection	121
Definition 4.1	Real-Time Bidding system	140
Definition 4.2	Inventory source	142
Definition 4.3	Publisher [Merriam-Webster.com, n.d.]	143
Definition 4.4	Supply Side Platform	143
Definition 4.5	Header bidding	144
Definition 4.6	Demand Side Platform	144
Definition 4.7	Bidding-as-a-service	145
Definition 4.8	Data Management Platform	145
Definition 4.9	Real-Time Bidding protocol	146
Definition 4.10	Private deal	148
Definition 4.11	Tag-Based Integration [AppNexus, 2017b]	149
Definition 4.12	RTB-bidding process [Merriam-Webster .com, n.d.]	150
Definition 4.13	Graph refactoring	157
Definition 4.14	Strict-consent mechanism	174
Definition 4.15	Implied-consent mechanism	175
Definition 4.16	Central node	187
Definition 4.17	Satellite node	187
Definition 4.18	Bridging node	187
Definition 4.19	Interconnecting node	188

List of figures

Figure 1.1	Big data and internet platforms	35
Figure 1.2	Visual structure of the thesis	44
Figure 3.1	Histogram (long-tail phenomenon)	95
Figure 3.2	Generic web-tracking model	106
Figure 3.3	Graph mining rules	107
Figure 3.4	RTB cookie 'uuid2'	112
Figure 3.5	Entropy versus counting third-party cookies	117
Figure 3.6	Example of a small-data model	121
Figure 3.7	Generic web-tracking model with Ghostery-node classification	125
Figure 3.8	Ghostery-node classification: missing values	126
Figure 4.1	Longitudinal analysis of the Rubicon Project partner network	138
Figure 4.2	Theoretical view of RTB systems	141
Figure 4.3	RTB sequence diagram with nine RTB-steps	151
Figure 4.4	EU heatmap and dendrogram	162
Figure 4.5	EU dendrogram	164
Figure 4.6	Edge boxplot of the subset top-20 edges	176
Figure 4.7	Edge boxplot of the full dataset	176
Figure 4.8	Cluster-edge betweenness: Slovenia (SI)	182
Figure 4.9	Central nodes: Slovenia (SI)	185
Figure 4.10	Annotated cluster: (Slovenia, SI)	186
Figure 4.11	Annotated cluster: Luxembourg (LU)	189
Figure D.1	Generic web-tracking graph: Denmark (DK)	267
Figure D.2	Partner network: the Rubicon Project (DK)	269
Figure D.3	Partner network: Adform (DK)	270
Figure D.4	Partner network: Adspine (DK)	271
Figure D.5	Partner network: AppNexus (DK)	272
Figure D.6	Partner network: Google (DK)	273

List of tables

Table 1.1	Structure of the thesis	44
Table 2.1	Overview of six WPM-research methodologies	54
Table 2.2	Common metadata for RTB client identification on desktop and mobile	64
Table 3.1	JavaScript and the HTML DOM	84
Table 3.2	Numeric summary and total number of . NL websites	93
Table 3.3	Bot detection report	101
Table 4.1	Analysis of distinct ad-slot sizes	142
Table 4.2	Overview of eight common RTB protocols	146
Table 4.3	EU presence of the Rubicon Project	168
Table 4.4	Luxembourg (LU): Four node types	188
Table B.1	Mapping of the filtered workshops with Research Questions (RQs)	258
Table C.1	Acquisitions and mergers (1/7)	259
Table C.2	Acquisitions and mergers (2/7)	260
Table C.3	Acquisitions and mergers (3/7)	261
Table C.4	Acquisitions and mergers (4/7)	262
Table C.5	Acquisitions and mergers (5/7)	263
Table C.6	Acquisitions and mergers (6/7)	264
Table C.7	Acquisitions and mergers (7/7)	265
Table D.1	Overview of news websites in Denmark (DK)	266

List of listings

Listing 2.1	Hotjar active fingerprint	65
Listing 2.2	Hotjar: ad-blocking detection	67
Listing 3.1	Simple script to select a crawl set from Alexa top1m (Python)	90
Listing 3.2	Command-line example parsing a mitm proxy data (Bash)	103
Listing 3.3	HTTP request header to a web resource	122
Listing 3.4	HTTP response header containing a HTTP 302 status code	122
Listing 3.5	Subgraph isomorphism: Matching pattern for a Cypher query	122
Listing 4.1	Example RTB bid request	147
Listing 4.2	Example of Netograph metadata	158

List of algorithms

Algorithm 3.1	Calculating the high-entropy value of a string s	116
---------------	--	-----

List of equations

Equation 3.1	Shannon entropy	114
Equation 4.1	Cluster-edge betweenness	181
Equation 4.2	Node betweenness	184

Data-driven ads

Today many websites show digital advertisements.¹ These ads seem to follow you when you are browsing from one website to the next one.² The tracking component of digital ads has given rise to serious privacy concerns.

Some end-users like the ads, others consider them as invading their personal privacy. However, what is privacy nowadays? The topic of the thesis is *understanding* the technology used to display these ads. This technology is called Real-Time Bidding (RTB). The key objective of this thesis is to answer the question to what extent can we measure the privacy component connected to RTB?³ To address the privacy component mentioned above, we will investigate a novel graph-based approach to analyze the complex data flows between a browser and a website.⁴

The present introductory chapter is outlined as follows. Section 1.1 introduces RTB. In Section 1.2 we provide our definition of web tracking and justify it by arguments (see Definition 1.8). By our definition of web tracking we position the thesis within the larger privacy debate. Section 1.3 serves as a background for our

¹ In the thesis we will colloquially call them ads.

² A simple experiment with a smartphone, a browser, and the Facebook app illustrates the tracking component. We retained two screen shots of (1) a webpage of a travel agency containing information about a city trip to Madrid, i.e., URL: <https://www.tui.nl/p/stedentrip/spanje/madrid/> (2 October 2017) and (2) moments later the sponsored ad by TUI for a city trip to Madrid in the Facebook-timeline.

³ On using the term component: I know very well that component is a notion with a large variety of interpretations. Of course, I have considered using the notion ‘aspect’, ‘facet’, or ‘element’, but in the end I always stumbled over the large number of *data* occurrences with respect to privacy. Therefore, I decided to maintain the notion ‘component’ just to stress its multi versatility.

⁴ We will also use the terms *user agent* and *host* for *browser* and *website*.

investigation into privacy and RTB systems. We will briefly look at data, big data, small data, and metadata. Section 1.4 presents the problem statement and formulates two research questions that the thesis aims to answer. In Section 1.5 we delineate the research methodology to answer the research questions. The chapter concludes by a description of the structure of the thesis in Section 1.6.

1.1 REAL-TIME BIDDING

Real-time bidding (RTB) is a recent phenomenon. No widely used academic definition exists and various terms are used. Throughout the thesis we use the term *real-time bidding* instead of the term *programmatic*.⁵ Based on the eight definitions provided in Subsection 1.3.1 and the areas of big data application introduced by OECD [2013a] (see Figure 1.1, p. 35), we propose the following definition for RTB.

DEFINITION 1.1: REAL-TIME BIDDING is a big data application within the organizational field of marketing to improve sales by *real-time* data-driven marketing and personalized (behavioral) advertising.

This definition matches various definitions used by advertising companies. IABureau [2012] and Duggan [2014] collected three of them,⁶ which we will reproduce here as an illustration of the stormy developments in the world of real-time advertisement.

DEFINITION 1.2: „REAL-TIME BIDDING is a software-based system to automate the buying, delivery, and optimization of ad campaigns.“ [Forrester]⁷

⁵ Laszlo and Smith [2015] differentiate between the two terms because according to them „programmatic represents a suite of markets which take advantage of technologies such as RTB auctions; but they also take advantage of buy transactions and in the future, the ability to trade futures.“

⁶ Throughout the thesis we refer to the Internet Architecture Board (IABoard) and the Interactive Advertising Bureau (IABureau). To avoid confusion, we will mark explicitly which organization we refer to. In this case IABureau [2012], i. e., Interactive Advertising Bureau (IABureau).

⁷ As cited by Duggan [2014].

DEFINITION 1.3: „REAL-TIME BIDDING means using technology and audience insights to automatically buy and run an ad campaign in real time, reaching the right person with the right message.” [Google]⁸

DEFINITION 1.4: „REAL-TIME BIDDING is a real-time system for either bidding on or buying ad inventory.” [IABureau, 2012] (shortened)⁹

Below we paraphrase from an interview by N. Singer [2012c]. Consumers generally remain in the dark as to how automated trading systems rank and therefore they shunt them. The systems allow companies to (1) segment,¹⁰ (2) prioritize, (3) target, and (4) re-target website visitors. Envision a Kafkaesque future where decisions are made about you and you do not know on what criteria they are based. RTB creates the possibility for companies to tag you wherever you are going, without you knowing or having the ability to influence it. It is becoming a huge imbalance for the ordinary end-user because, in the end, the ordinary end-user is the product.¹¹

RTB enables companies to process data on large scale, i. e., (personal) information related to an end-user about the websites he visits. The information may result in knowledge, i. e., an intrusive picture of one’s behavior and preferences and may qualify as data of a sensitive nature.

1.2 PRIVACY AND WEB TRACKING

The philosophical debate about the notion of privacy continues to take place since Warren and Brandeis [1890] who put *the right*

⁸ Ibid.

⁹ IABureau [2012], i. e., Interactive Advertising Bureau (IABureau), defines ad inventory as follows: „The aggregate number of opportunities near publisher content to display advertisement to visitors.” The pages of both terms were lastly modified on 4 April 2012. The full definition on the wiki contains the following additional information: „The initial RTB ecosystems evolved from the efforts of Demand Side Platforms (DSPs) to create a more efficient exchange of inventory. Due to these roots, RTB ecosystems put significant emphasis on end-user information (demographic and behavioral data, for example), while discounting the situation information (the publisher and context).”

¹⁰ Visitor segments can be created on customer data, e. g., age, estimated income, buying intent, household information.

¹¹ The paragraph is a paraphrase of a citation taken from an interview (with me) by N. Singer [2012c].

to be left alone square in its center as a reaction to the adoption of photography as a new technology.¹² A well-known cartoon published by the New Yorker (cf. K. Davis [2012, 17–18]) at a time when internet as a new technology was adopted headed: „On the internet, nobody knows you’re a dog”. Obviously, the meaning of the cartoon is more comprehensive in the age of big-data technology and the technical capability of end-user profiling. Resistance to new technology is determined partly by culture.¹³ For instance, in Asian cultures people are traditionally wary of being photographed. In Islamic cultures there is resistance against religious images. In contrast, in this thesis we work with the notion of informational privacy as defined by Westin [1967].¹⁴

DEFINITION 1.5: „PRIVACY is the claim of individuals, groups, or institutions to determine for themselves when, how and to what extent information about them is communicated.” [Westin, 1967, p. 7]

We apply this notion to the context of big data as a contemporary new technology, i. e., control over what (types of) data companies collect about people and how they use the data.

In addition to Westin’s privacy definition, he stated that „the individual’s desire for privacy is never absolute, since participation in society is an equally powerful desire.” In today’s context of big data, the fundamental right to privacy exists alongside other fundamental rights, e. g., :

- (1) freedom of expression, religion, and assembly/association,
- (2) non-discrimination, and
- (3) presumption of innocence/right of defence.¹⁵

The data protection principles of today’s legal privacy framework were already established in 1980 by the Organisation for Economic Co-Operation and Development (OECD) guidelines on the protection of privacy and transborder flows of personal data [OECD, 1980; 2013b]. Furthermore, we remark a second principle-based building block in data quality which was formulated by

¹² The notion is also known as *privacy as limited access*.

¹³ See Van Eijk [2016b, note 39, p. 159], addition by R. Hogendoorn (personal correspondence 15 September 2014).

¹⁴ See also, e. g., Zuiderveen Borgesius [2014, 87–92].

¹⁵ See, e. g., Blok [2018].

the convention of the Council of Europe for the protection of individuals with regard to automatic processing of personal data [Council of Europe, 1980, ETS No. 108] (which was modernized on 18 May 2018 [Council of Europe, 2018, CETS No. 223]). The eight OECD principles are on:

- (1) collection limitation,
- (2) data quality,
- (3) purpose specification,
- (4) use limitation,
- (5) security safeguards,
- (6) openness,
- (7) individual participation, and
- (8) accountability.

From the above, we remark that all seven fundamental principles in Article 5 of the General Data Protection Regulation (EU) 2016/679 (hereinafter: GDPR), including accountability, align with the OECD principles:

- (1) lawfulness, fairness, and transparency,
- (2) purpose limitation,
- (3) data minimization,
- (4) accuracy,
- (5) storage limitation,
- (6) integrity and confidentiality, and
- (7) accountability.

These (similar) principles are also enshrined in two Human Rights declarations and one Civil and Political Rights declaration:

- (A) Article 12 of the Universal Declaration of Human Rights,¹⁶
- (B) Article 17 of the International Covenant on Civil and Political Rights,¹⁷

¹⁶ Article 12: „No one shall be subjected to arbitrary interference with his privacy, family, home or correspondence, nor to attacks upon his honour and reputation. Everyone has the right to the protection of the law against such interference or attacks.”

¹⁷ Article 17: „(1) No one shall be subjected to arbitrary or unlawful interference with his privacy, family, home or correspondence, nor to unlawful attacks on his honour and reputation. (2) Everyone has the right to the protection of the law against such interference or attacks.”

- (c) Article 8 of the European Convention on Human Rights (ECHR).¹⁸

Furthermore, we mention five foundations for the Right to Privacy in Europe in which these principles are enshrined.

- (1) Article 16 of the Treaty on the Functioning of the European Union (TFEU),¹⁹ which provides the legal basis for the adoption of Union legal instruments relating to the protection of personal data.
- (2) Article 7 and Article 8 of the Charter of Fundamental Rights of the European Union (CFREU).²⁰ The charter has a similar legal value as the TFEU.²¹
- (3) General Data Protection Regulation (EU) 2016/679 (GDPR) [Parliament of the EU and the Council, 2016] repealing General Data Protection Directive 95/46/EC (GDPD) [Parliament of the EU and the Council, 1995]. The application and interpretation of the legal norms in the GDPR conform to the CFREU.²²

18 Article 8, right to respect for private and family life: „(1) Everyone has the right to respect for his private and family life, his home and his correspondence, (2) There shall be no interference by a public authority with the exercise of this right except such as is in accordance with the law and is necessary in a democratic society in the interests of national security, public safety or the economic well-being of the country, for the prevention of disorder or crime, for the protection of health or morals, or for the protection of the rights and freedoms of others.”

19 Article 16: „(1) Everyone has the right to the protection of personal data concerning them, (2) the European Parliament and the Council, acting in accordance with the ordinary legislative procedure, shall lay down the rules relating to the protection of individuals with regard to the processing of personal data by Union institutions, bodies, offices and agencies, and by the Member States when carrying out activities which fall within the scope of Union law, and the rules relating to the free movement of such data. Compliance with these rules shall be subject to the control of independent authorities. The rules adopted on the basis of this Article shall be without prejudice to the specific rules laid down in Article 39 of the Treaty on European Union.”

20 Article 7: „Everyone has the right to respect for his or her private and family life, home and communications.” Article 8 sees to the protection of personal data: „(1) Everyone has the right to the protection of personal data concerning him or her, (2) such data must be processed fairly for specified purposes and on the basis of the consent of the person concerned or some other legitimate basis laid down by law. Everyone has the right of access to data which has been collected concerning him or her, and the right to have it rectified, (3) compliance with these rules shall be subject to control by an independent authority.”

21 See also, Recital 1 GDPR.

22 In case of corresponding legal rights between the GDPR and the CFREU, the CFREU-rights have the same content as ECHR-rights as explained by the European Court

- (4) Regulation 45/2001/EC on the protection of individuals with regard to the processing of personal data by the Community institutions and bodies and on the free movement of such data [Parliament of the EU and the Council, 2000b].
- (5) Directive (EU) 2016/680 on the protection of natural persons with regard to the processing of personal data by competent authorities for the purposes of the prevention, investigation, detection or prosecution of criminal offences or the execution of criminal penalties, and on the free movement of such data, and repealing Council Framework Decision 2008/977/JHA [Parliament of the EU and the Council, 2000a].

With Westin's quote in mind, we identified eight categories of data in the context of big data, i. e., (1) the person's health, (2) economic situation, (3) information on political beliefs, (4) philosophical beliefs, (5) performance at work, (6) leisure, (7) personal preferences or interests, and (8) detailed geolocation or movements. These data categories indicate an individual's (sensitive) boundaries and fit our first attempt to define web tracking.²³

DEFINITION 1.6: „WEB TRACKING is the non-consensual collection, or processing, or storing of data, for the purpose of systematic monitoring, or profiling, of an end-user's habits across websites.“ [Van Eijk, 2011b, p. 19]

The definition requires a bird's eye view to be positioned in the ongoing philosophical debate about the meaning of privacy and the values it protects.²⁴ Many scholars have contributed to the

of Human Rights (ECTHR). See Article 52(3) CFREU on Scope and interpretation of rights and principles: „In so far as this Charter contains rights which correspond to rights guaranteed by the Convention for the Protection of Human Rights and Fundamental Freedoms, the meaning and scope of those rights shall be the same as those laid down by the said Convention. This provision shall not prevent Union law providing more extensive protection.“

²³ See our specialized Definition 1.8, which we will use throughout the thesis.

²⁴ We remark that Article 9(1) GDPR contains special categories of data „data revealing racial or ethnic origin, political opinions, religious or philosophical beliefs, or trade union membership, and the processing of genetic data, biometric data for the purpose of uniquely identifying a natural person, data concerning health or data concerning a natural person's sex life or sexual orientation.“ Furthermore, we remark Article 14 ECHR on Prohibition of discrimination: „The enjoyment of the rights and freedoms set forth in this Convention shall be secured without discrimination on any ground such as sex, race, colour, language, religion, political

theoretical conceptualization of privacy. We cite six of them as an illustration of numerous facets of the debate, i. e.,

- (1) liberty and security [Swire, Morell, Stone, Sunstein, & Clarke, 2013];
- (2) liberty and privacy [Schermer, 2007, pp. 71–90];
- (3) the distinction between the three concepts (a) privacy, (b) liberty, and (c) security [Klitou, 2012, pp. 15–18].
- (4) a proposal for a privacy-requirements ontology that distinguishes between privacy concerns regarding (a) significance of information, (b) expectations of harms, and (c) informational self-determination [Gürses, 2010];
- (5) the concept of harm impacting the private life of an individual or his family [Gratton, 2013]; and
- (6) the distinction between the three concepts (a) private sphere, (b) bodily integrity, and (c) informational privacy [Dubbeld, 2004, pp. 49–62].

The wider perspective enables us to connect the Web-tracking debate to the preceding privacy debate on Online Behavioral Advertising.²⁵ In addition to this connection, web tracking is feeding into the debates on (1) intelligence programs and big data, and on (2) social innovation and big data.²⁶ Social innovations are new ideas (products, services, and models) that simultaneously meet social needs (more effectively than alternatives) and create new social relationships or collaborations (cf. Murray, Caulier-Grice, and Mulgan [2010]).

To complete the wider picture, we note that the two overarching privacy debates serve to find answers to the following normative question.

*How to strike the right balance between (A), (B), and (C),
with
A: the security of a state and the protection of its civilians,*

or other opinion, national or social origin, association with a national minority, property, birth or other status."

²⁵ We consider Online Behavioral Advertising as a form of marketing (Figure 1.1, 35). See, e.g., Christl and Spiekermann [2016], Boerman, Kruikemeier, and Zuiderveen Borgesius [2017], FTC [2007; 2009], NTIA [2010], Article 29 Working Party [2010, WP 171], and Kotler and Armstrong [2012, pp. 519–520].

²⁶ See, e.g., Van den Herik and Van Eijk [2013; 2014; 2015; 2016]. Podesta, Pritzker, Moniz, Holdren, and Zients [2014] and Peppet [2014] address the difficulty of meaningful consumer consent in this context.

B: a (legitimate) interest for companies and its customers to benefit from big data, and
C: the (fundamental right to) privacy and freedom of people?

Now that we addressed the underlying privacy question, we will mark the start of the Web-tracking debate. It begins with Gomez, Pinnick, and Soltani [2009], Krishnamurthy and Wills [2009b], and Krishnamurthy and Wills [2010]. Relevant stakeholders convened at the Internet Privacy Workshop co-organized in December 2010 by the Massachusetts Institute of Technology (MIT), Internet Society (ISOC), World Wide Web Consortium (W3C), and the Internet Architecture Board. The workshop was a response from the standards community to the call from the Federal Trade Commission (FTC) on industry to create a Do Not Track (DNT) mechanism,²⁷ i. e., a mechanism that is easy to find and use, is effective and enforceable, and allows consumers to limit the collection of nearly all behavioral data gathered across sites and not just the serving of targeted ads (FTC [2010, pp. 66–69]; FTC [2013]).

After two workshops in December 2010 and April 2011,²⁸ almost all stakeholders joined the W3C Tracking Protection Working Group.²⁹ The working group hammered out a definition of *tracking*.³⁰ Through a painstaking process, the following definition was adopted.³¹

DEFINITION 1.7: „TRACKING is the collection of data regarding a particular end-user’s activity across multiple distinct contexts

²⁷ As described by Soghoian [2011], the DNT header field was based on the original DNT submission by Mayer, Narayanan, and Stamm [2011].

²⁸ See the workshop reports of *Internet Privacy Workshop* [2010] and *Web Tracking and User Privacy* [2011].

²⁹ See, e. g., McDonald [2018], Swire [2013], or Grimmelmann [2012] for a brief history of DNT and the formation of the W3C Tracking Protection Working Group.

³⁰ The W3C Tracking Protection Working Group’s decision policy includes the instrument of a Call For Objections (CFO). „If two or more competing proposals exist for an issue and the workgroup’s chairs conclude that further discussion on the proposals will not change existing positions, the chairs may conduct an electronic straw poll to call for objections to each of the presented proposals. Participants should express their objections to each proposal with clear and specific reasoning.”, URL: <http://www.w3.org/2011/tracking-protection/decision-policy.html> (4 April 2014).

³¹ See Doty and Mulligan [2013] for a detailed view of the multi-stakeholder process. The results of the CFOs for W3C-definitions for tracking and context respectively can be found here: URL: <https://www.w3.org/2002/09/wbs/49311/twpg-tracking-5/results> (5 April 2014), and URL: <https://www.w3.org/2002/09/wbs/49311/tpwg-context-240/results> (5 April 2014).

and the retention, use, or sharing of data derived from that activity outside the context in which it occurred. A context is a set of resources that are controlled by the same party or jointly controlled by a set of parties.” [Fielding & Singer, 2018]

During the course of the debate, Van Eijk proposed a second definition for *web tracking* by taking into account the element of a Unique Identifier (UID). This definition was presented at the IAPP Europe Data Protection Congress 2012 [Wefers Bettink, Van Eijk, & Wagner, 2012].³² We will use this specialized definition throughout the thesis. It is an important improvement of Definition 1.6.

DEFINITION 1.8: WEB TRACKING is an act by a party, or host, or service, of reading or writing Unique Identifiers (UIDs) that are connected directly or indirectly to an end-user, computer, or device while the end-user is interacting with various services of the web, in order to collect, combine, or analyze data about the end-user for charitable, philanthropic, or commercial purposes.

This definition includes various forms of market research,³³ e. g.,

- (A) outreach measurement (the degree to which end-users are served with ads across the web),
- (B) engagement measurement (the degree to which end-users interact with web services), and
- (C) audience measurement (the degree to which micro profiles can be derived from end-users interacting with services across the web).

Our specialized Definition 1.8 does not conflict with the definition of profiling in the General Data Protection Regulation (GDPR) [Parliament of the EU and the Council, 2016]. The GDPR-definition for profiling is as follows.³⁴

³² Cited by IWGDPT [2013, p. 1] and Zuiderveen Borgesius [2014, p. 53].

³³ Cf. IABureau [2017c, p. 12]: Measurement enables marketers to „compare metrics for an audience exposed to an ad, campaign, or channel, to the metrics for an audience not exposed to the ad, campaign, or channel. When the exposed group outperforms the control group, the difference in performance can be described as an incremental lift providing key insight for the marketers. The incremental results can be used to calibrate rule-based or statistical models to make sure that the attribution model reflects incrementality instead of just correlation.”

³⁴ An academic definition was proposed by Hildebrandt [2008]: „Profiling is the process of discovering correlations between data in databases that can be used

DEFINITION 1.9: „PROFILING means any form of automated processing of personal data consisting of the use of personal data to evaluate certain personal aspects relating to a natural person, in particular to analyse or predict aspects concerning that natural person’s performance at work, economic situation, health, personal preferences, interests, reliability, behaviour, location or movements.” [Parliament of the EU and the Council, 2016, Article 4(4)]

Below, we cite Recital 30 of the GDPR which clarifies the term profiling.³⁵

RECITAL 30: „Natural persons *may be associated with online identifiers* provided by their devices, applications, tools and protocols, such as internet protocol addresses, cookie identifiers or other identifiers such as radio frequency identification tags. *This may leave traces* which, in particular when combined with unique identifiers and other information received by the servers, *may be used to create profiles of the natural persons and identify them.*” [Parliament of the EU and the Council, 2016] (emphasis added)

However, profiling is also denoted - in Recital 24 - in the broader context of automated decisions. We cite Recital 24 of the GDPR, below.

RECITAL 24: „The processing of personal data of data subjects who are in the Union by a controller or processor not established in the Union should also be subject to this Regulation when it is related to the monitoring of the behaviour of such data subjects in so far as their behaviour takes place within the Union. In order to determine whether a processing activity can be considered to monitor the behaviour of data subjects, it should be ascertained whether natural persons are tracked on the internet *including potential subsequent*

to identify and represent a human or nonhuman subject (individual or group) and/or the application of profiles (sets of correlated data) to individuate and represent a subject or to identify a subject as a member of a group or category.”

³⁵ See also Council of Europe [2010]: „profiling techniques can enable the generation of new sensitive data concerning an identified or identifiable person; further considering that such profiling can expose individuals to particularly high risks of discrimination and attacks on their personal rights and dignity.”

use of personal data processing techniques which consist of profiling a natural person, particularly in order to take decisions concerning her or him or for analysing or predicting her or his personal preferences, behaviours and attitudes.” [Parliament of the EU and the Council, 2016] (emphasis added)

Recital 24 takes both the collection and the use of tracking data into account whereas Recital 30 and our Definition 1.8 primarily see to the collection of tracking data. Our specialized definition does take further use of tracking data into account, but only to the extent where it sees to processing aimed at the creation of profiles. It does not include further use into account, i. e., the application of the profile to the individual in order to take decisions concerning him.³⁶ Obviously, presenting a targeted advertisement based on a (micro) profile of end-user behavior is an important part of RTB. Simply put, RTB is a process (see Subsection 4.3.9). By using our own specialized definition of web tracking throughout the thesis we emphasize the data collection phase in RTB.

Concluding this section, we note that two elements are important for today’s notion of privacy: (1) transparency by disclosing data practices and (2) end-user control on, e. g., when data is collected, how long the data is retained, how data is used, and with whom the data is shared. *Transparency* and *end-user control* are the main issues guiding the current thinking on privacy. However, due to rapid advances in web tracking technologies, the two elements are undermined: (a) due to an *information asymmetry* to the detriment of the end-user and (b) due to the massive *collection of (personal) data* using big data technology.³⁷

1.3 DATA

The topic of this section concerns *data*.³⁸ In the thesis we use the following formal and widely accepted definition.

DEFINITION 1.10: „DATA are (raw) observations.” [Meesters, 2014]

³⁶ For brevity, we use ‘he’ and ‘his’ whenever ‘he’ or ‘she’ and ‘his or her’ are meant.

³⁷ See, e. g., Calo [2014]; Richards and King [2014]; Zuiderveen Borgesius [2014]

³⁸ Inter-University Consortium for Political and Social Research [n.d.]: „For social science, data is generally numeric files originating from social research methodologies or administrative records, from which statistics are produced.”

Next to data, the thesis deals with *information* and *knowledge*. Below we briefly describe the difference of these two concepts with data. We refrain, however, from formal definitions and a deep discussion since our emphasis is on data.

Information comes from the Latin verb *informare*. The best translation we provide here is forming an idea. The term is often used as follows: (1) transmission of information or (2) communication between a transmitter and a receiver.³⁹ Information is neither the same as knowledge nor as data; it is closely related with these two concepts. Data are (raw) observations, information is (scrubbed/pre-processed) data to which *meaning* is given.⁴⁰ Knowledge is data to which meaning is given and put into a *context*, basically it is information put into a context.

Definition 1.10 allows us to investigate three types of data.⁴¹ In the Subsections 1.3.1 to 1.3.3 we present the terms *big data*, *small data*, and *metadata*, and then turn to other key terms used in the thesis.

1.3.1 *Big data*

The term *big data* is relatively new. It is therefore not a surprise that at present, there is no widely used academic definition of big data.

As an introduction to the development in the field we illustrate the shift in meaning by no less than eight definitions. They are ordered chronologically. Such an aim for completeness (eight definitions) is not representative to every new term in this thesis, but here it serves the additional goal to show that the world of big data is a world of rapid and big changes. Lynch [2008] was one of the first to coin the term *big data*. He described the term as follows.

³⁹ In the digital domain the client-server model is the standard model for transmission/communication: the client sends a request and the server returns a response. See, e.g., Benatallah, Casati, and Toumani [2004].

⁴⁰ An internet poll from KDD-Nuggets shows that almost two-thirds of the respondents spend more than 60% of their time in a data-mining project on data cleaning and data preparation [Theus & Urbanek, 2008, p. 2]. URL: http://www.kdnuggets.com/polls/2003/data_preparation.htm (3 January 2015).

⁴¹ See Meesters [2014, pp. 83–85] for a hierarchy and the meaning of the nine concepts: (1) raw data, (2) scrubbed data, (3) information, (4) knowledge, (5) intuition, (6) norms, (7) values, (8) intelligence, and (9) action.

DEFINITION 1.11: „BIG DATA is data that will challenge the state of the art in computing, networking, and data storage.” [Lynch, 2008]

Lynch [2008] based his view on projects that generated huge amounts of data such as the Large Hadron Collider at CERN near Geneva, Switzerland. Subsequently, Jacobs [2009] was one of the first to propose a meta definition based on the notion of size. Then, Loukides [2010] noted that other data-centric companies, such as oil companies and telecommunication companies, already had big datasets for a long time and were used to working with data at scale. By combining Jacobs [2009] and Loukides [2010], we note the difference.

DEFINITION 1.12: „BIG DATA is data of which the size forces us to look beyond the tried-and-true methods that are prevalent at the time.” [Jacobs, 2009, p. 44] (slightly modified)

DEFINITION 1.13: „BIG DATA is when the size of the data itself becomes part of the problem.” [Loukides, 2010, p. 5]

After the first decade of the century, some people started to understand that big data was not only characterized by size (volume). For instance, Russom [2011, p. 6] identified three related attributes in an attempt to pin down big data in the meta-definition discussion. These attributes are: (1) volume, (2) velocity, and (3) variety. Beyer and Douglas [2012] and T. White [2012] have picked up these attributes, and attempted to formulate an academic definition.

DEFINITION 1.14: „BIG DATA is high-volume, high-velocity and high-variety information assets that demand cost-effective, innovative forms of information processing for enhanced insight and decision making.” [Beyer & Douglas, 2012]

It resulted in a nowadays widely accepted definition of big data by T. White [2012]. However, we note that his definition has a technical/engineering background. It is rooted in two new ideas: (1) novel ways to store data in a graph database and (2) new approaches to process the data.

DEFINITION 1.15: „BIG DATA is the term for a collection of datasets so large and complex that it becomes difficult to process using on-hand databases management tools or traditional data processing applications.” [T. White, 2012]

In terms of „data telling the story”, Loukides [2010] addressed tomorrow’s big data of narrative science and visualization models already then quite well. „Visualization is crucial to each stage of the data scientist” [Loukides, 2010, p. 7]. Only then, researchers could ask new questions that could not be answered a few years ago, due to insufficient computing power.

A key strategy that enabled researchers to answer new questions, was storing the information needed to „tell the story”, i. e., storing connected data or data about data,⁴² together with the data itself in a graph database. This new programming model allowed researchers to process data at unprecedented scale using distributed computing on computer clusters.⁴³

The challenges of the new programming model include [Van den Herik & Van Eijk, 2013; 2014]: (1) retention, (2) maintenance, (3) search, (4) exchange, (5) analysis, (6) visualization, (7) interpretation, and (8) real-time.

As a direct consequence of our remark on the engineering character of White’s definition, we show two definitions that attribute to public organizations, private organizations, and public-private interactions.

DEFINITION 1.16: „BIG DATA refers to things one can do at a large scale that cannot be done at a smaller one, to extract new insights or create new forms of value, into ways that change markets, organizations, the relationship between citizens and governments, and more.” [Mayer-Schönberger & Cukier, 2013]

⁴² We will look at the difference between the terms *data* and *metadata* in Subsection 1.3.3.

⁴³ See Robinson, Webber, and Eifrem [2015, pp. 193–210] and Patra [2015] for an overview on NOSQL. Patra [2015] groups NOSQL databases into four groups, each denoting a specific application, i. e., (1) columnar databases, e. g., Cassandra, HBase, (2) key-value store databases, e. g., DynamoDB, (3) document-store databases, e. g., MongoDB, CouchDB, and (4) graph databases, e. g., Neo4j, OrientDB. See also Harris [2013] for a brief history on Apache HADOOP, an open-source software framework to store and process data at scale.

DEFINITION 1.17: „BIG DATA refers to a value-creating focus to extract operational advantage from data and to enact it to deliver performance improvements. Two capabilities stand out, (1) learning faster from data and (2) putting that knowledge into action faster than competitors.” [Phillips, 2015]

Both definitions (1.16 and 1.17) refer to the question what can be done with big data. By combining Mayer-Schönberger and Cukier [2013, p. 6] and Phillips [2015] we can address the *application of data*. A recent report from the OECD [2013a] serves as an illustration. The report introduced five areas of a big data application within an organization. These five areas are as follows.

- (1) INNOVATION. Big data enables new products, goods, and services by making use of data, either as an independent product, or as a main component of a product that turns into a data-intensive product.
- (2) RESEARCH AND DEVELOPMENT. Big data provides a clear improvement of research and development capabilities.
- (3) ORGANIZATION. Big data encourages the development of a new organization structure and management thereof.
- (4) OPERATIONS. Big data supports optimization of production processes and logistic processes.
- (5) MARKETING. Big data improves sales by data-driven marketing and personalized (behavioral) advertising.

A well-known visualization to categorize the complexity of digital marketing is Brinker [2014].⁴⁴ The infographic categorizes 947 different companies into 43 categories across six classes. Below, we describe these classes.

- (1) MARKETING EXPERIENCES. The technology enabling the customer experience, interacting directly with the end-user.
- (2) MARKETING OPERATIONS. The information systems designed to improve sales by data-driven marketing and personalized (behavioral) advertising.

⁴⁴ Other popular visualizations are, e.g., CM Summit and BattelleMedia [2013], Lumapartners [2014], and Improve Digital [2015]. Improve Digital also compiles market maps for individual countries, e.g., 2017 Display Advertising Ecosystem for The Netherlands [Improve Digital, 2017e], Germany [Improve Digital, 2017c], Italy [Improve Digital, 2017d], Belgium [Improve Digital, 2017a], Spain [Improve Digital, 2017g], France [Improve Digital, 2017b], and Scandinavia [Improve Digital, 2017f].

- (3) MIDDLEWARE. The technology to manage the (big) data and its life cycle.
- (4) INTERNET. The marketing environment consisting of seven main platforms (a) email, (b) search, (c) video, (d) desktop, (e) mobile, (f) online games, and (g) social networks.⁴⁵
- (5) INFRASTRUCTURE. The technology to store (big) databases, and develop (big) data applications.
- (6) BACKBONE PLATFORMS. Integration components to automate and integrate marketing systems.

The relation to big data and other internet platforms, e. g., mobile or social networks, is shown in Figure 1.1. The seven internet platforms are well-known marketing *channels*. We will investigate the *big data* concerns of *display marketing* from an end-user perspective.⁴⁶

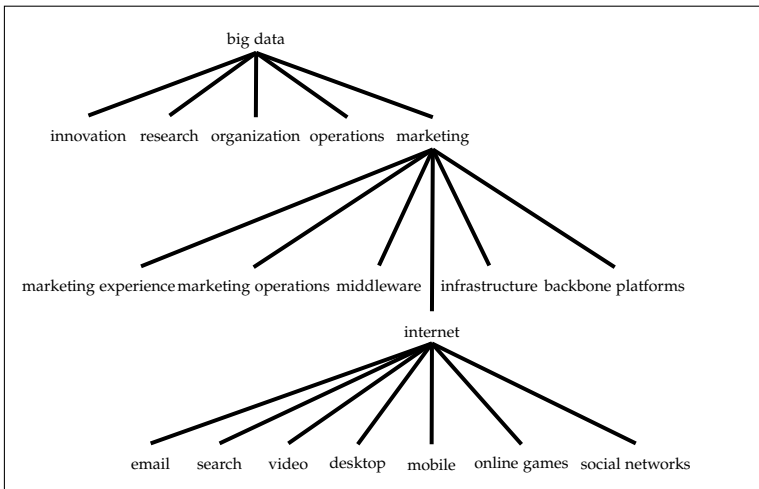


Figure 1.1: Big data and internet platforms. In the thesis we focus on internet applications of big data, i. e., desktop browsers and mobile browsers.

In passing we remark that the thesis entails data from interactions with services on the *internet* from private organizations and

⁴⁵ We did not include (smart) television and radio as internet platforms in our list.

⁴⁶ We also use the term display advertising throughout the thesis.

public organizations. Public organizations also use tracking technologies for various purposes, e. g., public broadcasting organizations with analytics solutions tracking their audience for various purposes, e. g., audience measurement.⁴⁷

We take the *desktop* as a starting point for our research.⁴⁸ However, the thesis does not deal with specific (additional) risks which may stem from the advent of the interactions on mobile devices (see, e. g., Leung, Ren, Choffnes, and Wilson [2016] and Ren, Lindorfer, Dubois, Rao, Choffnes, and Vallina-Rodriguez [2018]). Nevertheless, the principles for the desktop should also be applied for mechanisms used in other internet platforms. Eubank, Melara, Perez-Botero, and Narayanan [2013] surveyed tracking on the mobile web. The authors found many similarities between trackers that were used on mobile and those used on desktop. They also noticed that some cookies were expanding when new websites were visited. The difference in, e. g., the mobile-app context stems from the fact that apps operate in separated environments which restricts sharing of data across apps. Moreover, an end-user carries the mobile device with him (see, e. g., Article 29 Working Party [2011, WP 185, p. 7]). In other words, *geolocation* acts as a proxy for „the end-user’s real-world activity, interests and intent” [Laszlo & Smith, 2015, p. 9] which is specific for the mobile platform.⁴⁹

1.3.2 *Small data*

Now that we have staged data-driven innovation and marketing as two areas of big data application, we may have a closer look at the technology enabling them. For the enabling technique we use the term *small data*.

DEFINITION 1.18: SMALL DATA are the techniques enabling big data.

⁴⁷ See our specialized Definition 1.8 on web tracking.

⁴⁸ Ofcom [2015, p. 9]: „The computer (laptop/desktop/netbook) is still the primary device for accessing online content, but the use of alternative devices has increased substantially over the years. The popularity of both smartphones and tablets over the past five years have driven this, with two thirds (66%) of adults now using the former, compared to 30% in 2010.”

⁴⁹ See Section 2.2 for a comparison between the identification of an end-user in a desktop context and a mobile-app context. See Subsection 3.3.5 for an example of Hypertext Transfer Protocol (HTTP) redirection for email.

Rosenstein, Collins, and De Luca [1993] noted that „any rigorous definition of small data should be a function of dimension.” Nowadays we understand that big data is powered by small data.

COOKIES: AN EXAMPLE OF SMALL DATA

An example of small data are the cookies present in the Hypertext Transfer Protocol (HTTP).⁵⁰ Cookies have found widespread usage on the Web. Barth [2011, RFC 6265] documents the current usage of cookies already rather detailed. The cookie was originally standardized in RFC 2109 [Kristol & Montulli, 1997, RFC 2109].

DEFINITION 1.19: „A COOKIE is the state information that passes between an origin server and user agent, and that gets stored by the user agent.”⁵¹ [Kristol & Montulli, 1997, RFC 2109]

Cookie technology was created to solve a significant problem that was widely addressed by researchers in the 1990s. In the early days of the Web, sophisticated web services needed to store a state (i.e., to maintain a memory state) to operate an electronic shopping cart, or an interactive map to find the nearest store, or a live chat functionality. For such a rich customer experience, websites relied on cookie technology to pass variables from one web page to the next one. Web pages were still static at that time, comparable with the pages in a book. Until the cookie was introduced in 1994, user agents were not able to retain state information due to the mostly stateless nature of the HTTP protocol.⁵²

Nowadays, cookies are used for a whole range of purposes (see, e.g., IWGDPT [2013, pp. 3–4] for a short history of monitoring technologies). An important purpose of today’s cookies is to recognize an end-user when he visits a website.⁵³ The need to recognize an end-user enables the measurement of use of a service

⁵⁰ See Section 2.5. The Internet Engineering Task Force (IETF) standardized the initial HTTP protocol version with Request For Comments (RFC) 2068 [Fielding, Gettys, Mogul, Frystyk, & Berners-Lee, 1997, RFC 2068].

⁵¹ Cf. Fielding and Singer [2018]. In addition to the text to note 4, we use in the thesis the following meaning for *end-user* and *user agent*. An **END-USER** is a natural person who is making or has made use of the Web. A **USER AGENT** is any of the various client programs capable of initiating HTTP requests, including (but not limited to) browsers, spiders (web-based robots), command-line tools, custom applications, and mobile apps.

⁵² Cf. Tschofenig and Van Eijk [2011, p. 1].

⁵³ We categorize cookies into two types: (1) HTTP cookies and (2) JavaScript cookies.

on the Web, which allows companies to make decisions about the improvement of the end-user experience [IWGDPT, 2013, p. 1]. Companies rely on these measurements for a variety of marketing purposes, such as personalized advertisement, ad-measurement, and ad-verification.⁵⁴

We close this subsection with a quote by Latour [2007]: „It is as if the inner workings of private worlds have been pried open because their inputs and outputs have become thoroughly traceable“. We will give our view on web tracking in Section 1.2, but before doing so, we will look at the term *metadata*.

1.3.3 *Metadata*

To provide an idea of the difference between data and metadata, we start this section with an example. Thereafter, we focus on the importance of metadata. We illustrate the importance in the domain of telephone numbers. Then we show the result of connecting many domains. Finally, we propose our own definition of metadata.

THE DIFFERENCE BETWEEN DATA AND METADATA

The contents of a book is the story which the author is communicating to his readers. The title, the name of the author, and the publishing house are metadata, together with the book's unique ISBN. The story is the data (*purists* would say: the story is in the data). To understand the difference between data and metadata, we can learn from Schneier [2013], a renowned computer scientist. He illustrated the difference on his blog as follows.

„Imagine you hired a detective to eavesdrop on someone. He might plant a bug in their office. He might tap their phone. He might open their mail. The result would be the details of that person's communications. That's the *data*. Now imagine you hired that same detective to surveil that person. The result would be details of what he did: where he went, who he talked to, what he looked at, what he purchased – how he spent

⁵⁴ See, e.g., the Joint Industry Committee for Web Standards (JICWEBS) who approves reporting standards and codes for digital media. URL: <http://www.abc.org.uk/Rules-Regulations/Reporting-Standards-JICWEBS/> (7 April 2014).

his day. That's all *metadata*." [Schneier, 2013] (emphasis added)

THE IMPORTANCE OF METADATA

Metadata are the building blocks for detailed profiling.⁵⁵ Jan Koum is the founder of WhatsApp. With a snappy quote, he explained in an interview in the context of the acquisition of WhatsApp by Facebook why metadata are such important building blocks: „The numbers on your phone are *your real life network*“.⁵⁶

Metadata is created and retained at an unprecedented scale, not only by commercial companies, but also by public organizations using commercial data (we refer to, e. g., Greenwald [2013], Snowden [2014], or Gallagher and Greenwald [2014]).⁵⁷ This collection is strikingly illustrated with a quote by Edward Binney, a former National Security Agency (NSA) official.⁵⁸ He explained in an interview how data becomes metadata as follows.

„Every domain, think of a domain as an activity, a specific type of activity, phone calls, banking, or another

55 European Commission [2017] breaks down the notion of electronic communications *data* into electronic communications *content* and electronic communications *metadata*. Electronic communications *content* means „the content exchanged by means of electronic communications services, such as text, voice, videos, images, and sound“. Electronic communications *data* means „data processed in an electronic communications network for the purposes of transmitting, distributing or exchanging electronic communications content; including data used to trace and identify the source and destination of a communication, data on the location of the device generated in the context of providing electronic communications services, and the date, time, duration and the type of communication“.

56 The interview reads: „Koum made it [WhatsApp] one of the first mobile apps to sync with a phone's contacts. After he got fed up with forgetting his Skype user name and password, he went through the painstaking process of phone-number normalization for WhatsApp, ditching logins and passwords to make his service as simple as sending an SMS. The numbers on your phone are „your real-life network,“ he says. URL: <http://www.forbes.com/sites/parmyolson/2014/03/04/inside-the-facebook-whatsapp-megadeal-the-courtship-the-secret-meetings-the-19-billion-poker-game/3/> (8 March 2014).

57 Gallagher and Greenwald [2014] contains an image from a classified document of the TURBINE program which shows various commercial identifiers, e. g., the unique Google advertising cookies (PREFID). URL: <https://prod01-cdn03.cdn.firstlook.org/wp-uploads/2014/03/selectors-1024x768.png>. (16 April 2014)

58 The Guardian released a full NSA inspector general report (ST-09-002, working draft) on email, bulk internet metadata, content collection and analysis under Stellar Wind. URL: <http://www.theguardian.com/world/interactive/2013/jun/27/nsa-inspector-general-report-document-data-collection> (8 March 2014).

domain. So, if you think of graphing *each domain*, and then *each graph*, turning it into a third dimension. The trick now is to map through all the domains in that third dimension, pulling together all the attributes that any individual has in every domain, so that now I can pull your entire life together from all those domains and map it out and show *your entire life over time*.” [Poitras, 2012, 1’:47"-2’:27"] (emphasis added)

Three common examples of domain activities are browsing the web (browser history), logging in an online service (login time stamp history), and the use of e-mail (email metadata). Binney argued that it is the interconnection of domains that leads to the entire picture of your whole life over time. In fact, the interconnection of domains corresponds to the possibility to link data.⁵⁹

DEFINITION FOR METADATA

I would like to propose the following definition for metadata.

DEFINITION 1.20: METADATA is data about data to accurately describe (1) the properties of the data and/or (2) the relationships of the data with other data.

For completeness we remark the difference between the (unique) identifier (UID) of an end-user,⁶⁰ and his identity.⁶¹

DEFINITION 1.21: „An IDENTIFIER is identity information that unambiguously distinguishes one end-user, computer, or device from another one in a given domain.” [ISO, 2011, 24760-1, par. 3.1.4] (slightly modified)

⁵⁹ See also Article 29 Working Party [2014b, WP 216], which explains that pseudonymous data cannot be equated to anonymous information as they continue to allow an individual data subject to be singled out and linkable across different datasets.

⁶⁰ We remark that the recent California Consumer Privacy Act of 2018 [Secretary of State, 2018, AB-375] contains a definition for a UID: „a persistent identifier that can be used to recognize a consumer, a family, or a device that is linked to a consumer or family, over time and across different services, including, but not limited to, a device identifier; an Internet Protocol address; cookies, beacons, pixel tags, mobile ad identifiers, or similar technology; customer number, unique pseudonym, or user alias; telephone numbers, or other forms of persistent or probabilistic identifiers that can be used to identify a particular consumer or device.”

⁶¹ The importance of distinguishing between the two terms was pointed out by Ranenberg [2017] on the Internet Privacy Engineering Network (IPEN) mailing list.

DEFINITION 1.22: „An IDENTITY is a set of attributes related to an end-user, computer, or device.” [ISO, 2011, 24760-1, par. 3.1.2] (slightly modified)

1.4 PROBLEM STATEMENT AND RESEARCH QUESTIONS

In the beginning of this chapter and in Section 1.2 we touched upon the importance of the (normative) question regarding the technology used for data-driven ads. Moreover, we stated already there our aim: understanding the technology used to display these ads (Real-Time Bidding). This leads us to the formulation of the following Problem Statement (PS).

PROBLEM STATEMENT: *To what extent can we measure the privacy component connected to Real-Time Bidding?*

From this problem statement, we derive two Research Questions (RQs) to guide our research. The RQs are similarly formulated, but differ in focus. Although we recognize three distinct types of data handling, i. e., (1) the collection of data, (2) the use of data (e. g., profiling),⁶² and (3) the sharing of data,⁶³ we restrict our research to data *collection*. We aim to examine the actual data *collection* practices as they happen on the internet.

For clarity, the phrase „actual data collection practices” needs further explanation. By this, I mean the result of a four-step approach: (a) collecting data by visiting web pages, (b) filtering data to reduce complexity, (c) modeling small data in a directed property graph, and (d) examining relations and emergent phenomena in the graph structure with network science algorithms. With this approach in mind we are now ready to formulate our two RQs.

RESEARCH QUESTION 1: How do we move from data collection to graph analysis?

⁶² See, e. g., a recent study by Degeling and Nierhoff [2018]. They reported on the „automated measuring and influencing of Bluekai’s interest profiling”.

⁶³ Cf. Article 29 Working Party [2013a] The letter from the Art. 29 WP on W3C’s public consultation on the W3C Last Call Working Draft, 14 July 2015, Tracking Compliance and Scope includes the three types of processing as follows: „In the EU, consent for the processing of personal data includes the activity of collection. Therefore, the Draft DNT Compliance Specification must include instructions of ‘Do Not Collect’, ‘Do Not Use’, and ‘Do Not Share’.” (internal footnote omitted)

RESEARCH QUESTION 2: What are the emerging characteristics of the graph that is fit for graph analysis?

The main contribution arising from the thesis is a novel graph-mining approach to measure privacy *components* connected to RTB. The goal of the research is to develop a normative framework that addresses these *components*.

1.5 RESEARCH METHODOLOGY

Our research is empirical and exploratory research. Of course, it contains legal and societal aspects too. It builds upon Van Eijk [2011b], who explored the presence of tracking technology on European media websites. A total of 819 European media websites were visited to collect the relevant data for our research. Furthermore, the data was analyzed with (1) a specialized filter and (2) a dataset of confirmed tracking domains.⁶⁴

Our research aims at contributing to the understanding of the technology used to display the ads in real-time. We can only answer the problem statement (to what extent can we measure the privacy component connected to Real-Time Bidding) by investigating the (big) data that comes with real ads. Therefore, we aim at developing a new methodological approach based on nowadays available technologies. Later we will call this approach GBMA, meaning Graph-Based Methodological Approach.

We will unlock the data hidden in the HTTP data flow such that we can make sense of RTB systems. By graphing the small data through different lenses,⁶⁵ we will (re)create the story of the ads. This will contribute to a detailed view on web tracking.

Our research methodology is a stepwise approach. The eight steps are as follows.

- (1) Workshop participation.
- (2) Filter the challenging problems put forward in the workshops in such a way that we can combine the problems resulting in distilling a PS and the corresponding RQs.

⁶⁴ The specialized software to process the data are kept in a GitHub repository. URL: <https://github.com/rvaneijk/TDS> (3 December 2015). The open-source code to generate the specialized filter can be found here: URL: <https://github.com/rvaneijk/ruleset-for-AdBlock> (3 December 2015).

⁶⁵ Viz. Annotated Graph Mining (AGM) is a type of data mining that annotates the nodes and edges of a graph with metadata (cf. NWO [2006]): „The analysis of such graphs comprises both analysis of the graph structure and of the annotations.”

- (3) Literature review.
- (4) Conduct experiments from an end-user perspective and retain the research data.
- (5) Development of a metadata model and storing the data into a graph.
- (6) Analysis of the research data with intelligent techniques.
- (7) Compilation of a normative framework.
- (8) Evaluation of the framework.

In the enumeration above we have announced step 1 and step 2 as non-traditional steps in our research methodology. By their special nature we have step 1 listed in Appendix A and step 2 in Appendix B.

The research methods for conducting the experiments (step 4) and modeling the collected research data (step 5) are based on automated visits of websites. This technique is known as web crawling (see Definition 3.1 and Definition 3.2). In Chapter 3 we present a novel Graph-Based Methodological Approach (GBMA) which describes how we move from data collection to graph analysis (step 6). In Chapter 4 we examine RQ2 by applying the GBMA to digital media. In Section 4.4 we apply the GBMA to digital media by using it as a contextual data source. The research method for an empirical view of RTB systems is explained in detail in Subsection 4.4.2 and Subsection 4.4.3.

A final guideline for me as researcher is to be prepared to partition the research questions into subquestions when necessary. Of course, such a partitioning will be announced and supported by arguments.

1.6 STRUCTURE OF THE THESIS

The thesis consists of five chapters.

- CHAPTER 1. Sets the background and terminology, contains the PS and two RQs, and describes the research methodology to answer the RQs.
- CHAPTER 2. Contains a literature review.
- CHAPTER 3. Examines RQ1.
- CHAPTER 4. Examines RQ2 and contains a normative discussion about the privacy implications of RTB.
- CHAPTER 5. Combines the answers to RQ1 and RQ2 and provides the research conclusions. Moreover, the PS answered.

Below, we visualize an overview of the precise places where we deal with the RQs (Table 1.1) and provide the structure of the thesis (Figure 1.2).

Table 1.1: Structure of the thesis correlating PS and RQs to chapters.

	CH. 1	CH. 2	CH. 3	CH. 4	CH. 5
PS	✓				✓
RQ1	✓		✓		✓
RQ2	✓			✓	✓

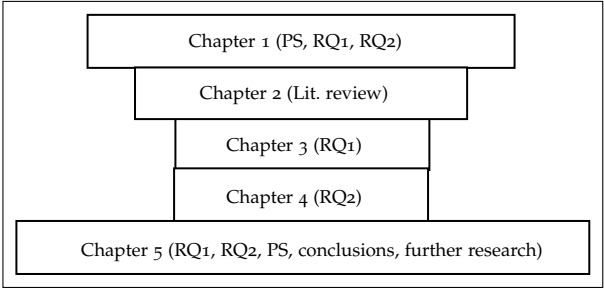


Figure 1.2: Visual structure of the thesis.

For now, I remark that so far, the Information Technology (IT) community has (virtually) developed a new technology. However, this does not mean that the members of the society have to give up their privacy. To gauge the situation, I will deeply investigate the matter and also consider what further research is to be recommended. Our journey will be continued in Chapter 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], Hiltz 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,0pt0ut.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, & Krusselbrink, 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 UUIDs 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 UUIDs 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 Big-Query [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).

```

1  hj.fingerprinter.prototype = { (...)
2      a = this.userAgentKey(a),
3      a = this.languageKey(a),
4      a = this.colorDepthKey(a),
5      a = this.timezoneOffsetKey(a),
6      a = this.sessionStorageKey(a),
7      a = this.localStorageKey(a),
8      a = this.indexedDbKey(a),
9      a = this.addBehaviorKey(a),
10     a = this.openDatabaseKey(a),
11     a = this.cpuClassKey(a),
12     a = this.platformKey(a),
13     a = this.doNotTrackKey(a),
14     a = this.pluginsKey(a),
15     a = this.adBlockKey(a),
16     a = this.hasLiedLanguagesKey(a),
17     a = this.hasLiedResolutionKey(a),
18     a = this.hasLiedOsKey(a),
19     a = this.hasLiedBrowserKey(a),
20     l = this.x64hash128(a.join("~~~"), 31));
21     return l
22 },      (...) };
23 hj.fingerprinter.VERSION = "0.7.1";
24 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., '("00000000" + (g[o] >> o).toString(16)).slice(-8) + ("00000000" + (g[1] >> o).toString(16)).slice(-8) + ("00000000" + (c[o] >> o).toString(16)).slice(-8) + ("00000000" + (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 Lindschow [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., 'o.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).

3

From data collection to graph analysis

This chapter addresses RQ1: how do we move from data collection to graph analysis? It means that we aim to find a path from collection to analysis. Therefore, we start at identifying methodologies for collecting research data. Each methodology is based on three key elements. They are the development of

- (1) software automating a browser in a way that it simulates an end-user visiting a website,
- (2) a proxy capturing and retaining HTTP data flows transmitted between the browser and the website, and
- (3) software managing the crawling movements of a large set of websites in a reliable way, e. g., by a queue manager.

The methodological basis including the three key elements has led to OpenWPM, the generic experimental framework (see Englehardt et al. [2014]).

To understand the real impact of the experimental framework on end-user privacy, we direct our (academic) focus on the transition path from *data collection* to *data analysis*. Therefore, we develop in this chapter a novel Graph-Based Methodological Approach (GBMA) to identify web tracking in HTTP header fields. Our approach consists of three parts.

PART 1. Data collection

PART 2. Data reduction

PART 3. Data modeling

Part 1 and Part 2 correspond with Step 4 in our research methodology (see Section 1.5), viz. to conduct experiments from an end-user perspective and retain the research data. Part 3 corresponds

with Step 5, viz. the development of a metadata model and storing the data into a graph. We discuss the three parts in the sections 3.1 to 3.3. We end this chapter with conclusions (Section 3.4) and we provide an answer to RQ1 (Section 3.5).

3.1 DATA COLLECTION

The first part of the GBMA is data collection. For this purpose, we conduct experiments by simulating an end-user's browser actions and retain the research data. Several key elements are weighed carefully before crawling the web. Below, we will discuss browser data collection (Subsection 3.1.1), data-flow collection (Subsection 3.1.2), and proxy collection (Subsection 3.1.4).

3.1.1 *Browser data collection*

In this subsection we restrict the scope of our perspective to interactions with services from the desktop. Basically, we take services that can be browsed from a desktop computer as a starting point.

Moreover, we restrict our research to the collection of research data from an end-user viewpoint. In doing so we consider the (desktop) browser as the instrument closest to the end-user.¹⁶¹

There are various ways of collecting research data from an end-user viewpoint. We identify three different ways of collecting research data, i. e., (A) browser centric, (B) server centric, and (C) network centric.

A: BROWSER CENTRIC

A browser centric approach is typically based on either *customizing* the browser itself (henceforth: a browser patch) or *extending* the browser with specialized software (henceforth: a browser extension). Below we provide examples to illustrate the difference between the two approaches.

¹⁶¹ This contrasts with browsing with a smart mobile device. For mobile one can argue that the device itself and the device identifier(s) connected to it are closest to the end-user. The fact that an end-user carries a smart mobile device with him makes it closer to the end-user than a desktop browser.

A1: BROWSER PATCH

Three examples of a specialized browser patches are Prividor [Schütte, 2011], FPDetective [Acar et al., 2013, p. 667], and Privaricator [Nikiforakis, Joosen, & Livshits, 2015]. Local changes were made to the source code, i. e., Prividor and FPDetective modified the Firefox browser to their needs, and Privaricator modified the Chromium browser which is the open-source version of Google's Chrome browser.

A2: BROWSER EXTENSION

Two examples of browser extensions to gather research data are Live HTTP Headers [Savard & Co, 2011],¹⁶² and FourthParty [Mayer & Mitchell, 2012].¹⁶³

Browser patches and browser extensions operate on the level of the browser DOM API and are accessible with JavaScript. The browser DOM API enables access to (1) the HTML DOM which contains cookies [Barth, 2011, RFC 6265] and (2) browser (pre-)loads of resources [Grigorik, 2015]. Below we focus on browser extensions.

FIVE BROWSER OBJECTS

The DOM API gives access to browser objects such as (1) window, (2) navigator, (3) screen, (4) location, and (5) document. We note that the browser object location is not the same as geolocation. It means the originating web server. In Table 3.1 we provide a brief description of the five browser objects. Three of them (window, navigator, and document) have properties.¹⁶⁴ The total number of properties is six (see Table 3.1, next page). Two of them (screen and location) do not have properties in our table.¹⁶⁵ The description illustrates the capabilities to process data through DOM APIs

¹⁶² See, e.g., Krishnamurthy and Wills [2009a] or Van Eijk [2011b] who used the Live HTTP Headers extension to collect their WPM-research data.

¹⁶³ [Mayer & Mitchell, 2012] presented the FourthParty web measurement platform. Others used the browser extension to collect research data, e.g., Eubank et al. [2013]. FourthParty inspired OpenWPM [Englehardt et al., 2014] and the creation of the browser extension 'Tracking Report Card' [Adams, Fedor, Felten, Felten, Fenton, Mayer, Ralli, Toubiana, Turner, & Van Eijk, 2012].

¹⁶⁴ Throughout the thesis we use Uniform Resource Identifier (URI) for a link on a webpage and Universal Resource Locator (URL) for the domain of a website.

¹⁶⁵ Properties of screen are, e.g., height, width. URL: https://www.w3schools.com/jsref/obj_screen.asp (19 February 2018). Properties of location are, e.g., host,

of modern browsers.¹⁶⁶ First of all, we remark that the spelling of ‘referrer’ in the browser object `document.referrer` originates from a misspelling in [Berners-Lee, Fielding, & Frystyk, 1996, RFC 1945, pp. 44–45]. Furthermore, we provide adequate references. The main publication is by Powers [2008], but also others deserve credits, e. g., Van Kesteren, Gregor, Ms2ger, Russel, and Berjon [2015, DOM4], Le Hors, Le Hégarret, Wood, Nicol, Robie, Champion, and Byrne [2004, DOM3 Core], Hickson, Berjon, Faulkner, Leithead, Doyle Navara, O’Connor, and Pfeiffer [2014, HTML5], WHATWG [2015, HTML Living Standard], and Berners-Lee, Fielding, and Masinter [2005, RFC 3986]).

Table 3.1: JavaScript and the HTML DOM.

BROWSER OBJECT	DESCRIPTION
<code>window</code>	Represents the open window in a browser.
<code>window.history</code>	The URI of the HTML pages visited by the end-user.
<code>window.name</code>	The property sets or returns the name of the window.
<code>navigator</code>	Information about the browser.
<code>navigator.userAgent</code>	The end-user agent string of the browser.
<code>screen</code>	Information about the screen of the device.
<code>location</code>	The URI of the current HTML page visited by the end-user.
<code>document</code>	The entire HTML page that has been loaded.
<code>document.URL</code>	The URI of the current HTML page visited by the end-user.
<code>document.cookie</code>	The cookies for the current HTML page.
<code>document.referrer</code>	The URI of the document that loaded the HTML.

Browser extensions are particularly useful for collecting data from (panel)users who opted-in to sharing their (personal) data. An advantage of the extensions is that it is rather easy to adopt

origin. URL: https://www.w3schools.com/jsref/obj_location.asp (19 February 2018).

¹⁶⁶ For instance, Janc and Zalewski [2014] argued that „it is possible to use `window.name` to store persistent identifiers for a given window: if an end-user deletes all client state but does not close a tab that at some point in the past displayed a site determined to track the browser, re-navigation to any participating domain will allow the window-bound token to be retrieved and the new session to be associated with the previously collected data.”

the system to new requirements of the end-user environment. For instance, in August 2015 Mozilla announced that they changed the browser extension API (called WebExtensions) of the Firefox browser. Originally, WebExtensions was designed to be compatible with the API model of the Chrome browser. The change made it easier for developers of add-ons to build extensions that will work across multiple browsers.¹⁶⁷

Thereafter it was easier to collect data and share it with others under predefined conditions due to the enhanced interoperability between different browsers. Many privacy-browser extensions provide this telemetry. TrackerBlock [Brock, 2012] - a Firefox browser extension - was one of the first in the field of WPM to collect aggregated tracking statistics from its end-users. Three more examples are the well-known extensions Ghostery [Pierce, 2010], Electronic Frontier Foundation (EFF)'s Privacy Badger [Eckersley et al., 2014], and Disconnect [Kennish, Chau, & Oppenheim, 2013].

A disadvantage of loading a browser extension is that it increases the entropy of the browser.¹⁶⁸ High browser entropy increases the risk identification by a (unique) browser fingerprint (cf. Eckersley [2010a], see also, e.g., Starov and Nikiforakis [2017a] or Haga et al. [2017, p. 1,665–1,668].).

Although fingerprinting cannot be prevented, a (partial) remedy to mitigate the risk is to use a default browser without browser extensions (see, e.g., Doty [2015]; Tillmann [2013]). The Tor-browser is an example of a browser that goes to great lengths in mitigating the risk [The Tor Project, 2015].

B: SERVER CENTRIC

A server centric approach requires *direct access* to the server's log files. Server data of this kind is usually hard to get access to. For instance, Yahoo released several large server centric datasets to the public in 2016.¹⁶⁹ Two of these are specifically interesting for RTB researchers, i.e., (a) Data Targeting End-User Modeling [Yahoo, 2014] and (b) Search Marketing Advertiser Bidding Data [Ya-

¹⁶⁷ Cf. Needham [2015].

¹⁶⁸ Supra Section 2.4.

¹⁶⁹ URL: <https://webscope.sandbox.yahoo.com/catalog.php?datatype=a> (30 April 2017).

hoo, 2002]. Both sets contain a sample of advertiser's bid and revenue information.¹⁷⁰

We mention three studies on server centric data: (1) Zhang, Yuan, Wang, and Shen [2014], (2) Budak, Goel, Rao, and Zervas [2014], and (3) Aziz and Telang [2015]. Zhang, Yuan, Wang, and Shen [2014] analyzed a dataset that contained the server-side data of three seasons of the iPinYou global RTB algorithm competition. Budak et al. [2014] analyzed web browsing histories of 14 million individuals. Aziz and Telang [2015] had access to and analyzed a dataset containing over 30 million bid requests for 586,909 unique individuals who are potential customers of an advertiser.¹⁷¹

Here we note, that there are usually almost 60 bid requests stored for each individual. The data is suited for studying end-user behavior from an algorithmic perspective. The set of bid requests contained three particular data properties (a) the amount that was bid for in every bid request, (b) whether the bid was successful, and (c) whether the potential customers made any purchase in the next three days.

C: NETWORK CENTRIC

Unlike the two approaches above, a network centric approach *captures HTTP data flows* transmitted between the browser and the website. The data is retained by specialized software such as Mitmproxy [Cortesi et al., 2010]. Mitmproxy is a proxy written in Python. The software is capable to intercept data transmitted through HTTP Secure (HTTPS).

For our approach Mitmproxy is the basis. This is also true for, e.g., OpenWPM [Englehardt et al., 2014] and MobileScope [Soltani et al., 2012]. Its data format is widely used and contributes to the sharing of data between researchers.

¹⁷⁰ Yahoo [2014] contains „a small sample of advertiser's bid and revenue information over a period of 4 months. Bid and revenue information is aggregated with a granularity of a day over advertiser account id, key phrase, and rank. Apart from bid and revenue, impressions and clicks information is also included." The data is a snapshot of end-user activity data collected at Yahoo servers during the period from 8 April 2014 to 20 July 2014. (cf. Wiley [2017]).

Yahoo [2002] contains „the bids over time of all advertisers participating in Yahoo Search Marketing auctions for the top 1000 search queries collected during the period from 15 June 2002 to 14 June 14 2003."

¹⁷¹ The origin of the iPinYou dataset is the iPinYou DSP.

OTHER FORMATS, E. G., HAR

There exists software parsers to convert the research data into other data formats, e. g., the widely used HTTP Archive (HAR) format [Odvariko, Jain, & Davies, 2012]. It is important to know that Nassri [2015] provides a list of resources that support HAR. At the time of writing, the list contained 27 applications.

HAR enables further analysis with, e. g., (1) browser-debugging tools for web developers, (2) specialized network-analysis software, or (3) network performance measurement tools such as BigCloud (see, e. g., Nassri [2015] or Grigorik [2013]; Kahle [1996]). BigCloud is a cloud service operated by Google. The service enables researchers to upload a HAR dataset into a Dremel database [Melnik, Gubarev, Long, Romer, Shivakumar, Tolton, & Vassilakis, 2010] and access their data via a Representational State Transfer (RESTful) API with SQL-like queries. The R-package Bigrquery [Citro, 2014] bridges BigCloud with R's statistical toolbox [R Core Team, 2014]).

The Mitmproxy data can easily be parsed into smaller subsets. This (1) enables prototyping and (2) makes everyday data juggling with big crawls considerably easier. For instance, a small set of research data - e. g., 10% of all the data records - is usually sufficient for prototyping and testing different data models.

At the end of this paragraph on network centric, we would like to communicate one lesson learned from a change in the Mitmproxy data format.¹⁷² The Mitmproxy change led to a problem in parsing a crawl from January 2013. The crawl was collected with Mitmproxy version 0.8 and a recent crawl 2 with version 11.2. As it turned out the results of the two crawls were no longer comparable. We solved the issue by parsing the historic dataset with a corresponding version of Mitmproxy.¹⁷³

3.1.2 *Data-flow collection: existing methods*

For data-flow collection (based on the network centric approach) we need specialized software. As a case in point, we mention the software testing framework Selenium [Selenium Project, 2004] that enables the collection of data flow by automating the Fire-

¹⁷² Cf. the change in the browser extension API.

¹⁷³ To revert to a specific version of Mitmproxy on Ubuntu we used the following command: `sudo apt-get install package=<version>`.

fox browser. In this environment Van Eijk and Terra [2012] automated website visits with a Python script. Before we discuss our approach to data-flow collection for our first experiments, we remark that other ways of gathering data flows exist. Below we give four examples based on specialized web development tools designed for web developers to test and debug their code: (A) Python scripting library, (B) PhantomJS, (C) debugging protocol of the Chrome browser, and (D) W3C Webdriver. Then we continue with (E) our 3-step procedure of data-flow collection for our *first experiments* of which the advantages are also described. We note that our procedure of data-flow collection for our *later experiment* for the purpose of our empirical view of RTB systems (see Section 4.4) is described in Subsection 4.4.2.

A: PYTHON SCRIPTING LIBRARY

Researchers and browser-extension developers can use a scripting library such as Python's Requests module (see, e.g., Lavrenovs and Melon [2018] or Reitz, Benfield, Cordasco, and Petrov [2011]).¹⁷⁴ Scripting with a module capable of making HTTP requests is a quick way of automating web visits to collect research data.¹⁷⁵ One important thing to keep in mind when experimenting with different scripting libraries is to take care of support for cookies. That is, whether the session and the persistent cookies are retained. Support for cookies is important for end-user impersonation, which we will discuss in more detail in Subsection 3.1.4.

B: PHANTOMJS

Many researchers use PhantomJS [Hidayat, 2011], a headless browser. The scriptable JavaScript API makes it fit for the purpose of data-flow collection and for WPM (see, e.g., Some, Rezk, and Bielova [2018] or Hannak, Sapiezynski, Molavi Kakhki, Krishnamurthy, Lazer, Mislove, and Wilson [2013]). PhantomJS is built on WebKit which is also the rendering engine for, e.g., the Safari browser [Apple, 2005]. Recently, the browser vendors Mozilla [Claypotch, 2017] and Google [Bidelman, 2017] announced that

¹⁷⁴ The source code of the module is available on GitHub. URL: <https://github.com/kennethreitz/requests> (15 December 2015).

¹⁷⁵ An example of a simple Python command to visit a web page is, e.g., `resp = requests.get('http://campusdenhaag.nl/crk')`.

their browsers are ready for headless usage. The announcement is good news for WPM scholars, since it enables them to use headless Firefox or Chrome as a substitute for PhantomJS. We remark that Linux already offered a workaround for data collection in ‘headless mode’ by using virtual framebuffers.

C: DEBUGGING PROTOCOL

Many browser-extension developers use the (remote) debugging protocol for the Chrome browser, which is part of the Chrome Developer Tools [Google, 2015]. The protocol is scriptable with Python. Clift [2014] deploys the protocol by sending JSON messages [Bray, 2014, RFC 7159] via a network connection to enable the automation of a browser instance. This approach is particularly interesting for WPM researchers who are looking to conduct experiments on a smart phone, tablet, or another mobile device. Finally, we remark that the early version of Netograph [Cortesi, 2011] was built on Mitmproxy [Cortesi et al., 2010]. Recently, the codebase of Netograph [Cortesi, 2017] has been brushed off and changed to the Webdriver debugging protocol [Stewart & Burns, 2018]. It has a main focus on (1) large scale continuous crawling, (2) storing processed metadata in Google Bigtable [Google, n.d.-a], and (3) providing API access for further analysis.¹⁷⁶

D: W3C WEBDRIVER

The Webdriver protocol enables a WPM researcher to automate site visits.¹⁷⁷ Similar to (1) Selenium [Selenium Project, 2004] and (2) the Chrome debugging protocol, W3C Webdriver [Stewart & Burns, 2018] is an attempt to standardize the control and introspection of browsers.

3.1.3 *Our 3-step procedure of data-flow collection*

Now that we discussed four different ways of automating the browser, we will deepen the understanding of our GBMA to collecting flow data (i.e., Part 1 of the GBMA). For our first exper-

¹⁷⁶ URL: <https://netograph.io/docs/api/docs> (27 August 2018).

¹⁷⁷ Stewart and Burns [2018]: „WebDriver is a remote control interface that enables introspection and control of user agents. It provides a platform- and language-neutral wire protocol as a way for out-of-process programs to remotely instruct the behavior of web browsers.”

iments we used the Alexa website. We are aware that many researchers use other datasets, e.g., the ranking of Quantcast¹⁷⁸ or SEOmoz.¹⁷⁹ Our choice was guided by the quality of the results of the crawls (see Table 3.2). We start discussing three collection steps (called C-steps). The three C-steps are performed in accordance with the experimental framework that is fit for long-tail experiments.

C-STEP 1: FILTERING ALL .NL WEBSITES

First, we defined our crawlset by filtering all .NL websites from the top1m dataset of that day.¹⁸⁰ We provide Listing 3.1 as an example of such a filter.

Listing 3.1: Simple Python script to select a crawlset from Alexa top1m [Van Eijk & Terra, 2012].

```

1 from urllib2 import urlopen
2 from lxml import html
3
4 for n in range(20):
5     d = html.fromstring(urlopen("http://www.alexa.com/
        topsites/countries;%d/NL" % (n, )).read())
6     for site in d.cssselect("div.desc-container h2 a"):
7         print site.get("href").split("/")[-1]
```

The script starts with a declaration of two helper modules (Listing 3.1, rr. 1–2). Because the top ranking .NL websites are not listed on one long page, but on 20 pages, we need to loop through the Alexa website to collect our selection (Listing 3.1, rr. 4–5). We used the HTML tag ‘div.desc-container’ as an easy way to parse the HTML page (Listing 3.1, r. 6). The last line makes sure the output is written in a pretty format (Listing 3.1, r. 7).

178 Quantcast provides another popular dataset that ranks websites based on the number of people in the Netherlands, who visit a website within a period of time. URL: <https://www.quantcast.com/top-sites/NL> (20 July 2015).

179 SEOmoz provides the Moz Top 500, a list of domains ranked by the number of linking root domains. URL: <https://moz.com/top500> (31 march 2018).

180 Top1m means the first one million websites of that day. URL: <http://www.alexa.com/topsites/countries/NL> (14 January 2016).

C-STEP 2: CONSTRUCTING A REPRESENTATIVE CRAWL

Second, we started to construct the crawl with

- (1) a fresh browser instance,
- (2) empty cache, and
- (3) an empty profile, i. e., neither browsing history nor cookies (henceforth: first view).

Each domain on the list of target websites was visited twice but not in the same order. We randomized the list before visiting the sites for the first time (henceforth: round 1), and randomized the list again before visiting the sites for the second time (henceforth: round 2). No links were clicked or followed. No ads were clicked. Moreover, we also did not log in to any (social media) services or click on any social media buttons. In later experiments we performed more crawls (up to six crawls).

We remark that websites usually do not publish a social widget on their homepage which makes detecting the trackers less effective with our type of crawl. To measure *tracking by social widgets*, a deep crawl method is needed. The issue of end-user impersonation will be discussed in Subsection 3.1.4.

C-STEP 3: CRAWLING A LARGE SET IN A RELIABLE WAY

Third, Van Eijk and Terra [2012] used a queue manager to manage crawling a large set of websites in a reliable way. The script is written in Python. A queue manager is necessary. The queue manager has three functions,

- (1) to load the set of websites, i. e., performing round 1 and then round 2, one by one,
- (2) to overcome browser crashes and (programming) problems restraining the browser to finish loading the entire document, and
- (3) to maintain the browser state by retaining the Firefox profile. The profile contains important end-user data, i. e., browsing history and cookies.

Without a queue manager the end-user data could be lost in the event of a browser crash or when a page-load problem occurs. To prevent this loss from happening, the queue manager retains

the Firefox profile and subsequently terminates the Firefox browser instance. Before launching a new Firefox instance, it restores the Firefox profile, such that the crawl can continue where it was interrupted.

We continue our research in five paragraphs, viz. (A) the first experimental results; (B) the results analyzed; (C) the long tail experimental results; (D) the impact of technical measures and (E) conclusions on data-flow collection.

A: THE FIRST EXPERIMENTAL RESULTS

For the crawl itself, we used a shallow crawl method instead of a deep crawl.¹⁸¹ Below we provide a definition for shallow crawl (Definition 3.1) and deep crawl (Definition 3.2).

DEFINITION 3.1: A SHALLOW CRAWL is a process of visiting the top domain of each website in a list of websites.

DEFINITION 3.2: A DEEP CRAWL is a process of visiting the top domain and/or visiting (a number of) internal links of each website in a list of websites.

We compiled our list of websites by selecting all the Dutch websites (.NL top level domain) from the Alexa [n.d.] top 1,000,000 list of websites in the world (henceforth: top1m dataset). For the shallow crawls (listed in Table 3.2) each domain name in the crawl list was visited without visiting (deep) links.

Table 3.2 shows the number of .NL websites in the top1m dataset for six different crawls. There we provide a numerical summary of the six crawls with the measures of meaning in the third column 'Numerical' and the corresponding result in the fourth column 'Summary'.

From the numerical data in Table 3.2 we may conclude that the average number of the most popular websites in the Netherlands is about 8,000 (to be precise, the mean value is 8,277) as considered with respect to the top1m popular websites in the world. To put these numbers into perspective, the number of registered .NL

¹⁸¹ See also, e.g., Hoofnagle and Good [2012a; 2015] for the difference between shallow and deep crawling.

websites in 2014 was over 5 million (to be precise: 5,531,186) according to the annual report of SIDN [2014], the Dutch top-level domain administrator.

Table 3.2: Numeric summary and total number of six different crawls of .NL websites in the Alexa top 1,000,000 list of popular websites.

CRAWL	.NL WEBSITES	NUMERICAL	SUMMARY
Crawl1 (Nov 2012)	8,472	Min.:	7,968
Crawl2 (Dec 2012)	8,067	1st Qu.:	8,094
Crawl3 (Jan 2013)	7,968	Median:	8,294
Crawl4 (Apr 2013)	8,415	Mean:	8,277
Crawl5 (Mar 2015)	8,174	3rd Qu.:	8,458
Crawl6 (Jun 2015)	8,567	Max.:	8,567

B: THE RESULTS ANALYZED

Despite a clear selection criterion to compile our crawl set, the total number of Dutch websites in Table 3.2 differs from other sources. The difference can be explained by looking closer at the Alexa’s ranking methodology (See Alexa [2015a]. Below we will briefly discuss (B1) Alexa’s ranking methodology and (B2) the importance of the long-tail phenomenon for WPM.

B1: ALEXA’S RANKING METHODOLOGY

Alexa collects data from panel members who opted in to the sharing of their browsing data. Alexa counts who makes actual visits to which websites in a rolling three month period. The ranking of a website is based on two elements, (1) the number of panelists who visit a website on a given day and (2) the number of page views by the panelists.

However, websites are ranked relatively to other websites. Therefore, changes in traffic to one website affect the ranking of others. This is also known as a long-tail phenomenon (cf. Alexa [2015b]), which we will discuss below (see B2 and C).

B2: THE IMPORTANCE OF THE LONG-TAIL PHENOMENON FOR WPM

Long-tail phenomena are well studied in other (academic) fields.¹⁸² However, this is not the case for WPM, although his conclusions of WPM studies are often based on data collected from rankings of the top-websites.¹⁸³ Therefore, it is - from a WPM-researcher's viewpoint - important to understand the long-tail phenomenon of ranking. It helps us to understand (1) what is happening when we crawl a set of websites and (2) whether different crawls can be used in longitudinal studies.

C: THE LONG TAIL EXPERIMENTAL RESULTS

In the generic experimental framework, the choice for using (parts of) datasets such as Quantcast's ranking and Alexa's top1m for crawling has been guided by a personal preference.¹⁸⁴ Table 3.2 shows six examples of crawls of .NL websites. The long-tail phenomenon for .NL websites is illustrated in Figure 3.1.

The histogram of Figure 3.1 presents (a) the number of .NL websites (topnl) in relation to the Alexa rank, (b) the percentage of .NL websites (percentnl) relative to their ranking, and (c) the percentage of .COM websites (percentcom) relative to their ranking. The top1m dataset of 22 June 2015 contained 486 .NL websites ranked under ranking 100,000. This number counts for about 5% of the .NL websites in the top1m dataset. The remaining 95% of .NL websites ranked 100,000 or lower and are subject to the long-tail phenomenon. More than half of the .NL websites rank very low, i. e., ranked 524,289 or lower.

Furthermore, Figure 3.1 shows that .NL websites are slightly (about 6%) more prone to the long-tail phenomenon than .COM websites. Websites with a low ranking, are much more sensitive

¹⁸² See also, e.g., Anderson [2007] for a detailed description of the long-tail phenomenon. Anderson [2007] examined the three questions: (1) „What happens when there is almost unlimited choice?“, (2) „What happens when everything becomes available to everyone?“, and (3) „When does the combined value of the millions of items that only sell in small quantities equal or even exceed the value of a handful of best-sellers?“.

¹⁸³ See also, e.g., Papadopoulos, Kourtellis, and Markatos [2018] or Scheitle, Jelten, Hohlfeld, Ciprian, and Carle [2018]: „The conclusions of *security* studies are often based on data collected from domains that appear on commercial rankings of the 'top' websites. However, the data sources and methodologies used to compile these rankings vary widely and their details are unknown, leading to hidden properties and biases in the rankings.“ (Scheitle et al. [2018, p. 13], emphasis added).

¹⁸⁴ Supra n. 178.

to small increases or decreases in traffic. The long-tail ranking applies to both .NL and .COM websites. Straightforwardly stated, websites ranked 100,000 or lower can be identified ranking in the long-tail of top1m websites.

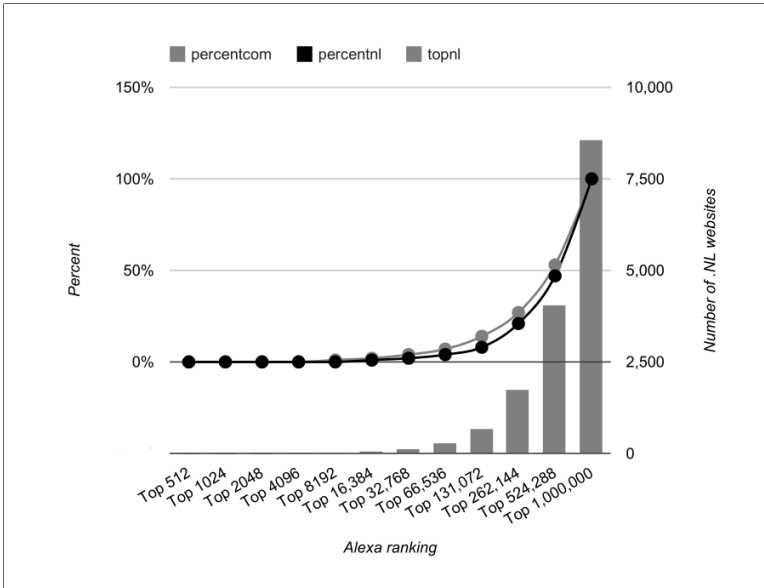


Figure 3.1: Histogram of (a) the number of .NL websites (topnl) in relation to the Alexa rank, (b) the percentage of .NL websites (percentnl), and (c) the percentage of .COM websites (percentcom). The percentages are relative to the total websites, i.e., respectively 8,567 .NL websites and 554,768 .COM websites. The Alexa global top1m dataset was retrieved 22 June 2015, from URL: <http://s3.amazonaws.com/alexa-static/top-1m.csv.zip>.

D: THE IMPACT OF TECHNICAL MEASURES

We note that data-flow collection will become harder as technical measures to mitigate the risk of State surveillance improve (see, e.g., Kranch and Bonneau [2015]). The aim of the risk measures is to mitigate a breach of confidentiality of the communication. Obviously, this has an impact on WPM.

There are two clear drawbacks to the measures to mitigate the risk: (1) the measures make it harder to collect research data through web crawls, and (2) the measures restrict the transparency and ability to verify the marketing technology. Three examples of the application of relevant technical measures to WPM are:

- (1) HTTP Strict Transport Security (HSTS) [Hodges, Jackson, & Barth, 2012, RFC 6797],
- (2) public key pinning extension for HTTP [Evans, Palmer, & Sleevi, 2015, RFC 7469], and
- (3) Transport Layer Security (TLS) Error Reports [Moeller & Langley, 2015, RFC 7507].

Below we give a brief characterization of these three measures.

- Ad 1. HSTS is a measure enabling a web site and a user agent to interact only over secure connections. It prevents piggyback on commercial tracking cookies to identify end-users.
- Ad 2. Public-key pinning is a measure to counter man-in-the-middle attacks due to compromised Certification Authorities.¹⁸⁵ Technically, Mitmproxy cannot intercept pinned data flows once the original certificates are retained in a browser.
- Ad 3. Firefox checks server configurations for insecure connections due to protocol downgrade attacks on depreciated versions of TLS.¹⁸⁶ This behavior may result in websites not loading during a crawl.

E: CONCLUSIONS ON DATA-FLOW COLLECTION

Data-flow collection will become harder as technical measures to mitigate the risk of state surveillance will be stricter. These developments have an impact on data-flow collection according to our GBMA approach. So, collecting data with the Chrome debugging protocol has the potential to become the default approach for collecting data for WPM researchers. In all, we have three conclusions.

¹⁸⁵ [Hodges et al., 2012, RFC 6797] defines a new HTTP header that allows „web host operators to instruct user agents to remember (‘pin’) the hosts’ cryptographic identities over a period of time. During that time, user agents will require that the host presents a certificate chain including at least one Subject Public Key Info structure whose fingerprint matches one of the pinned fingerprints for that host.”

¹⁸⁶ Firefox message: Error code: ssl_error_unsupported_version URL: <https://support.mozilla.org/en-US/kb/tls-error-reports> (10 July 2015).

CONCLUSION 1: In conjunction with Table 3.2 we may conclude that filtering .NL websites from the top1m dataset on any given day with the help of the GBMA results in a crawl set fit for studying web tracking despite the long-tail effect.

CONCLUSION 2: Because .COM websites account for the overall majority of websites ranked in the daily top1m dataset, we may conclude that the long-tail phenomenon to take into account is a maximum of 6% of the researcher's crawl set.

CONCLUSION 3: If the study is longitudinal, then datasets over time will not be linkable in its entirety assuming that the domain name is the key to link the datasets.

3.1.4 *Proxy collection*

The goal of proxy collection is end-user impersonation. Our approach to automate the browser was conducted from an end-user viewpoint with the latest version of Firefox at the time of crawling. According to our methodology for the collection of research data we used three key elements (See the introduction of this chapter, viz. (1) automating a browser (2) using a proxy and (3) using a queue manager). The browser was installed to run headless on a minimal installation of the latest version of the Ubuntu server. The data was collected in the Netherlands over a standard consumer Asymmetric Digital Subscriber Line (ADSL)-connection to the internet.

For technical reasons, we made two minimal changes to its default configuration. First, we installed a self-signed certificate to enable the interception of HTTPs because Mitmproxy cannot intercept the protocol without it.¹⁸⁷ Second, for the same reason, we disabled support for SPDY, which is a network protocol that uses header compression. HTTP/2 (Peon and Ruellan [2015, RFC 7541], Belshe, Peon, and Thomson [2015, RFC 7540]) is replacing SPDY. Yet, the configuration of header compression in Firefox is located in the same browser-configuration subkeys.¹⁸⁸ This subsection consists of two parts, viz. (A) end-user impersonation and (B) data-management issues on the data collection.

¹⁸⁷ Cf. Cortesi et al. [2010].

¹⁸⁸ In Firefox, i. e., 'about:config', the keys to configure HTTP header compression are located under 'network.http.spdy.enabled.http2'.

A: END-USER IMPERSONATION

We took in our experiments the following seven elements into account: (A1) end-user events, (A2) end-user profiling, (A3) localization, (A4) fingerprinting, (A5) bot detection, (A6) end-user consent, and (A7) forensic screenshots. These elements are closely related to the online behavioral patterns of the end-user. In effect, we see that behavior is often repeated. This leads to diversification into distinct patterns, e. g., visiting favorite websites in a specific order around the same time each day. For now, we are only interested in the phenomenon of proxy collection. Therefore, we disregard a classification that is too detailed. Below we give a straightforward characterization of the seven elements A1 to A7.

A1: END-USER EVENTS

Each end-user event is straightforward. No JavaScript emulation of mouse movement and/or scrolling is used to refine the impersonation of an end-user. Many complex end-user interactions can be emulated by the Selenium Project [2004] end-user facing API.¹⁸⁹ An example of end-user interaction required by default by the Firefox browser is ‘click to play’. The feature is a recent security measure to reduce malware infections based on (known) vulnerabilities of the browser add-ons, e. g., the Adobe Flash player. Technically, the browser inserts a wrapper that requires end-user interaction before the content is passed to a browser add-on.¹⁹⁰

A2: END-USER PROFILING

End-user profiling is straightforward. No (pre-)training of a (browser) profile is used to impersonate an end-user by visiting specific categories of websites.¹⁹¹ We crawled with an empty browser state. In contrast, Van Eijk [2011b] visited Dutch children’s websites three days prior to collecting research data, which resulted

189 I.e., the JavaScript Object ‘org.openqa.selenium.interactions.Actions’. Selenium gives a researcher many options to simulate end-user interactions, e.g., a mouse wiggle. See URL: <http://selenium.googlecode.com/git/docs/api/java/org/openqa/selenium/interactions/Actions.html> (15 January 2016).

190 Firefox versions 8 and above support click-to-play. In Firefox, i.e., in Firefox’s configuration settings (‘about:config’) the key to configure click-to-play is located under ‘plugins.click_to_play’.

191 Supra Definition 1.9.

in an advertisement for a Donald Duck subscription in Dutch on a German news website.

A big challenge with this approach is to keep the profiles clean. The risk of polluting the profile accidentally is real which means that it is difficult to maintain a *chain of evidence* (a term used in computer forensics related to the methodology of managing evidence that is destined for use in court).

A3: LOCALIZATION

Although a localized Dutch version of the Firefox browser exists, i.e., 'firefox-locale-nl', we installed the default Firefox browser. In addition, no changes were made to the user agent string or the referrer header.¹⁹² The referrer header is an important field in the HTTP request header [Fielding & Reschke, 2014b, RFC 7231, Section 5] providing information about the request context.¹⁹³

A4: FINGERPRINTING

Except from refraining from browser extensions, e.g., the Flash plug-in that allows end-users to play streaming media, no measures to mitigate browser fingerprinting were taken. An example of a fingerprinting mitigation measure is the automated generation of random browser settings while crawling. We refer to Gulyas, Some, Bielova, and Castelluccia [2018], a recent study on the uniqueness of browsers resulting from the presence of browser extensions.¹⁹⁴

The technology landscape for streaming media is changing. The Flash technology is being replaced by HTML5 due to its greater interoperability capabilities. From a technical perspective, a Flash cookie is not that different from a HTTP cookie. Even though Flash cookies are stored in a different place on the end-user's computer, Flash technology transmits and receives cookies over HTTP. This means that Flash cookies can be collected by a network centric approach as well as with a browser centric approach.

¹⁹² The referrer header can be disabled by changing the Firefox configuration setting 'Network.http.sendRefererHeader'. The default value is '3'. Changing it to a value of '0' causes the Firefox browser to omit a referrer field in the HTTP request header [Fielding & Reschke, 2014b, RFC 7231, Section 5].

¹⁹³ Supra Section 2.5.

¹⁹⁴ Gulyas et al. [2018] analyzed how unique users are based on their behavior. They reported that almost 55% (54.86% to be precise) of the 16,393 user in their dataset that have installed at least one detectable extension are uniquely identifiable.

In fact, Google announced in February 2016 that it is phasing out Flash ads.¹⁹⁵ Google's strategy is allow only display ads in HTML5. Moreover, Mozilla removed support for flash.¹⁹⁶ Oracle's strategy for sites running Java technology was to encourage its implementers to migrate to the Java Network Launching Protocol (JNLP).¹⁹⁷ It is quite in time now that Adobe announced on 25 July 2017 that it will end support for Flash by the end of 2020.¹⁹⁸ Other companies, including Microsoft,¹⁹⁹ Apple,²⁰⁰ and Facebook,²⁰¹ have made similar announcements in response to the changing streaming media landscape.

A5: BOT DETECTION

We must assume that various threat mitigation measures have been put into place to deal with ad fraud. Yet, we did not implement a bot-detection mitigation strategy.

Threat mitigation measures by RTB actors are necessary since publishers and website visitors have been subject to injected malware in ad slots on a web page.²⁰² For instance, Majumdar, Kulkarini, and Ravishankar [2007] investigated click fraud in Content Delivery Networks (CDNs). A CDN is a network provider, e. g., Akamai, a company that facilitates the worldwide distribution of ads. Majumdar et al. [2007] proposed four protocols to mitigate replay and fabrication attacks for CDN providers. We provide two exam-

195 URL: <http://doubleclickadvertisers.blogspot.nl/2016/02/google-display-ads-go-100-html5.html> (25 July 2016).

196 Mozilla announced in 2015 to remove support for most plugins for (streaming) media content, e. g., Microsoft's Silverlight and Oracle's Java by the end of 2016. URL: <https://blog.mozilla.org/futurereleases/2015/10/08/npapi-plugins-in-firefox/> (10 October 2015).

197 URL: http://docs.oracle.com/javase/8/docs/technotes/guides/deploy/applet_dev_guide.html#JSDPG1032 (10 October 2015).

198 URL: <https://blogs.adobe.com/conversations/2017/07/adobe-flash-update.html> (26 July 2017).

199 URL: <https://blogs.windows.com/msedgedev/2017/07/25/flash-on-windows-timeline/#D6TjufFSsIgJ03TZ.97> (26 July 2017).

200 URL: <https://webkit.org/blog/7839/adobe-announces-flash-distribution-and-updates-to-end/> (26 July 2017).

201 URL: <https://developers.facebook.com/blog/post/2017/07/25/Games-Migration-to-Open-Web-Standards/> (26 July 2017).

202 An example of a popular Dutch website that was affected by 'malvertising' is nu.nl. URL: https://www.security.nl/posting/395932/Nu_nl+laat+advertenties+extra+op+malware+controleren (12 December 2015).

ples of specialized software capable of detecting link spam caused by bots crawling websites.²⁰³

First, we mention Bad Behavior [Jaquith, Firas, Error, & Skeltoac, 2014] which analyzes the HTTP headers, IP address, and other metadata regarding HTTP requests to determine bot activity. For instance, if the software detects non-compliance with the HTTP State Management Mechanism [Kristol & Montulli, 2000, RFC 2965], it may ignore the bot and block its traffic. Table 3.3 contains a (partial) Bad behavior bot detection report illustrating common bot detection strategies. The data was captured from a website with a PHP implementation of Bad Behavior.²⁰⁴

Table 3.3: Bad Behavior bot-detection report.

REASON	COUNT	PERCENT
Required HTTP header 'Accept' missing	40	75.47%
Web browser attempted to send a traceback	11	20.75%
False claim. User agent claimed to be Googlebot	1	1.89%
User agent claimed to be MSIE with invalid OS version	1	1.89%
Banned proxy server in use	0	0%
HTTP header 'Connection' contains invalid values	0	0%
HTTP header 'Referer' present but blank	0	0%
IP address found on (external) blacklist	0	0%
Required HTTP header 'Accept-Encoding' missing	0	0%
Rotating user agents detected	0	0%
Prohibited HTTP header 'Range' present	0	0%
Prohibited HTTP header 'via' present	0	0%
Prohibited HTTP header 'Proxy-Connection' present	0	0%
User agent was found on blacklist	0	0%
User agent is required but none was provided	0	0%
Use of rotating proxy servers detected	0	0%

²⁰³ See also the Interactive Advertising Bureau (IABureau)/ABC Spiders & Robots Lists [Audit Bureau of Circulations, 2013]. Other examples are, e.g., HostKarma, Spamhaus, SORBS, Google Safe Browsing, and Yandex Safe Browsing.

²⁰⁴ The bot-detection report was retrieved on 6 July 2015 from our own dokuwiki webserver URL: <http://www.blaeu.com>. To save costs, the domain has been reconfigured to serve email only since 20 July 2015. The bot-detection report has been retained in our lab log on 6 July 2015.

Second, we mention the Project Honey Pot [Unspam Technologies, n.d.] which is, in contrast to Bad Behavior, a distributed spam detection solution. Project Honey Pot provides a completely automated public Turing test to tell computers and humans apart (CAPTCHA) [Rosenberg & Jennings, 2008, RFC 5039]. Recently, CloudFlare - a cloud provider - started using Project Honey Pot to protect their clients from spambots.²⁰⁵

A6: END-USER CONSENT

For the end-user consent the following holds. Neither explicit consent for tracking cookies was expressed, e.g., by agreeing to a cookie statement, nor implied consent was given, e.g., by clicking three or more deep links.

A7: FORENSIC SCREENSHOTS

We did not capture a screenshot of the webpage after rendering, although the Selenium WebDriver includes this functionality. However, for certain types of experiment, e.g., checking compliance with the Dutch Cookie law, the information is relevant. The forensic screenshot serves as evidence of the information provided to the end-user at the time of crawling the webpage.

B: DATA-MANAGEMENT ISSUES ON THE DATA COLLECTION

Above, the necessary description of the elements used in our experiments is given. Other data management issues are stored in accordance with the rules given by the Leiden University and in particular by the Faculty of Law. This ends the first step in our Graph-Based Methodological Approach (GBMA).

3.2 DATA REDUCTION

The second part in our GBMA is data reduction. There, the key objective of the methodology is to retain precisely the research data that we need for our data model. To better understand this

²⁰⁵ See CloudFlare's FAQ. URL: <https://support.cloudflare.com/hc/en-us/articles/203366080-Why-do-I-see-a-captcha-or-challenge-page-Attention-Required-trying-to-visit-a-site-protected-by-CloudFlare-as-a-site-visitor-> (10 July 2015).

part, we embark upon the difference between the two terms data and information. Data are (raw) observations, i.e., in our case the Mitmproxy data. Information is pre-processed data to which meaning is given.²⁰⁶ Reducing the research data requires three reduction substeps (called R-steps).

R-STEP 1: Parse the Mitmproxy data.

R-STEP 2: Ensure the data quality.

R-STEP 3: Retain the small data in a graph.

Below we discuss the three R-steps. We guide the discussion by a basic example of the extraction of HTTP response header fields (see Listing 3.2).

Listing 3.2: Command-line example parsing a Mitmproxy data file (Bash).

```

1  $ cat parse.py
2  def response(context, flow):
3  ext = tldextract.extract(flow.request.host)
4      request_headers = [{"name": k, "value": v} for k, v in flow
5                          .request.headers]
6      response_headers = [{"name": k, "value": v} for k, v in
7                          flow.response.headers]
8      print response_headers
9
10 $ mitmdump -n -q -s parse.py -r <file>.mitm | \
11 $ cut -d : -f 1 | sort | uniq > parsed_headers.csv

```

R-STEP 1: In the first R-step, we parsed the Mitmproxy data with an inline script. This script is based on the 'flow' object included in the 'libmproxy' Python library.²⁰⁷ We used a Python script (Listing 3.2, r. 2–6) to extract the HTTP response headers and piped commands (Listing 3.2, r. 8–9), i.e., the commands are separated by the character '|', to retain the research data. Below we briefly explain the four commands.

- (1) Parse the crawl to extract all HTTP header fields.
- (2) Filter (see 'cut -d : -f 1' which parses out column 1 from a colon (:) delimited input).

²⁰⁶ Cf. Meesters [2014, pp. 83–85], supra Section 1.3.

²⁰⁷ We refer to Cortesi, Shaver, and Wingert [2014] for more examples of inline scripting.

- (3) Sort.
- (4) De-duplication (see 'uniq') of all HTTP header fields.

The outcome is a file (`parsed_headers.csv`) containing a list of distinct HTTP response header fields.

The parser interfaces with the research data through Mitmproxy's API. The flow object contains methods to parse information (a) from the HTTP request header and (b) from the HTTP response header. With Python we extracted typical crawl variables, e.g., referrer, location, user agent, and cookie.²⁰⁸ We remark that the cookies are shaped as tuples by Mitmproxy. A tuple is a sequence of immutable Python objects.²⁰⁹

R-STEP 2: In the second R-step (after the required data is extracted), the data is ready for inspection for incorrect or incomplete values. Browser crashes or page-load problems that occurred during the crawling process may have impacted the quality of the small data.

R-STEP 3: In the third R-step we imported the small data into a data model that is appropriate for graph-based analysis. In our research we use Neo4j, an open-source graph database with access to the research data via a RESTful API [Robinson et al., 2015]. Neo4j comes with a number of language drivers, e.g., Go, Java(Script), .NET, Perl, PHP, Python, R, and Ruby. We experimented with Perl, Python and R bindings. The RNeo4j R package [N. White, 2014] connects Neo4j with R's statistical toolbox which includes the iGraph R package.²¹⁰

CONCLUSION ON DATA REDUCTION

In conclusion, data reduction is a necessary part of GBMA. The aim is selecting small data from collected raw data required for data modeling.

²⁰⁸ See also the respective RFC, i.e., 'Referer' [Fielding & Reschke, 2014b, RFC 7231, Section 5.5.2], Location [Fielding & Reschke, 2014b, RFC 7231, Section 7.1.2], User agent [Fielding & Reschke, 2014b, RFC 7231, Section 5.5.3], and Cookie [Barth, 2011, RFC 6265].

²⁰⁹ URL: http://www.tutorialspoint.com/Python/Python_tuples.htm (15 August 2015).

²¹⁰ See also Csárdi and Nepusz [2014]; R Core Team [2014]

3.3 DATA MODELING

The third part in our GBMA is the development of a *small-data* model. The key objective of the small-data model is to unlock the information defining web tracking within the *big* dataset itself. Below, we will discuss the following notions: a generic web-tracking model (Subsection 3.3.1), the formalization of graph-mining rules (Subsection 3.3.2), an extended web-tracking model (Subsection 3.3.3), high-entropy identifiers (Subsection 3.3.4), and a small-data model (Subsection 3.3.5).

3.3.1 *Generic web-tracking model*

The idea of a generic representation of the (third-party) web tracking problem (henceforth: generic web-tracking model) was first proposed by Van Eijk [2011b].²¹¹ Nowadays, other researchers use different terms, e. g., cross-domain tracking [Schneider, Enzmann, & Stopczynski, 2014, pp. 23–24], cookie matching [Mazel, Garnier, & Fukuda, 2017, p. 3], or cookie syncing (see, e. g., Papadopoulos et al. [2018], Acar et al. [2014, p. 682], Englehardt et al. [2015, pp. 6–7], Englehardt and Narayanan [2016, p. 11], Bashir, Arshad, Robertson, and Wilson [2016, pp. 491–493], and Brookman et al. [2017, p. 137]).²¹²

Figure 3.2 (next page) illustrates the generic web-tracking model as an undirected graph. If a tracker (C) receives information from websites (A) and (B) about the same end-user, then (C) may track the end-user across different websites by linking the information. As Barth [2011, RFC 6265] puts this eloquently:²¹³

„Although cookies are not the only mechanism, that servers used to commonly track end-users across HTTP requests, cookies facilitate tracking because they are persistent across user-agent sessions and can be shared between hosts.”

²¹¹ Supra our specialized definition for web tracking (Definition 1.8).

²¹² Example of cookie syncing: I captured the UID ‘6254484537578825505’ being synced between Criteo and AppNexus on the webpage <http://t2tea.com> (16 July 2018): <https://dis.criteo.com/dis/rtb/appnexus/cookieismatch.aspx?appnxsid=6254484537578825505>.

²¹³ Barth is also known for defining the concept Same Origin Policy (SOP) [Barth, 2011, RFC 6265]. The SOP is implemented in modern browsers, e. g., as a security measure to protect e. g., HTTP cookies against the risk of access to the cookie data by JavaScript(s).

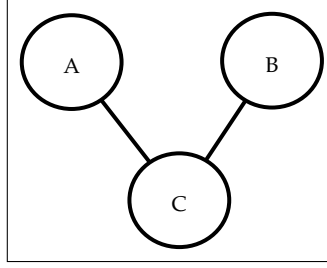


Figure 3.2: Generic web-tracking model [Van Eijk, 2011b].

In *this* thesis we start afresh and formalize the generic web-tracking model. Let us start adopting Definition 3.3 of a labeled property graph as proposed by Riesen, Jiang, and Bunke [2010].

DEFINITION 3.3: „GRAPH. Let L_V and L_E be a finite or infinite label alphabet for nodes and edges, respectively. A graph g is a four-tuple $g = (V, E, \mu, \nu)$, where V is the finite set of nodes, $E \subseteq V \times V$ is the set of edges, $\mu: V \rightarrow L_V$ is the node labeling function, and $\nu: E \rightarrow L_E$ is the edge labeling function.” [Riesen et al., 2010, pp. 219–220].

If we combine Van Eijk [2011b] and Definition 3.3, the formal description of web tracking is as follows.

DEFINITION 3.4: WEB TRACKING. Let us consider a graph denoted by $g = (V, E, \mu, \nu)$ with three nodes (V_A, V_B, V_C), two edges ($E_1: V_A \times V_C, E_2: V_B \times V_C$), and three edge properties ($\mu_A=A, \mu_B=B, \mu_C=C$). If a tracker (C) receives information from websites (A) and (B) about the same end-user, then (C) may track the end-user across different websites by linking the information.

3.3.2 Formalization of graph-mining rules

Van Eijk [2011b] diligently investigated small data to detect their importance for web tracking.²¹⁴ In this subsection we will reflect

²¹⁴ The source code of the specialized HTTP header-analysis software was written in the PERL programming language. The code and the results of 148 different experiments are available in a GitHub repository. URL: <https://github.com/rvaneijk/TDS/> (3 December 2015).

on the findings and formalize the results with respect to Definition 3.4. Formalization of the graph patterns (Figure 3.3) allows us to reason within a system. We discuss (A) illustration and formalization of small data, (B) induced subgraphs and its applications, (C) motif enumeration and counting algorithms, and (D) conclusions on the formalization of graph-mining rules.

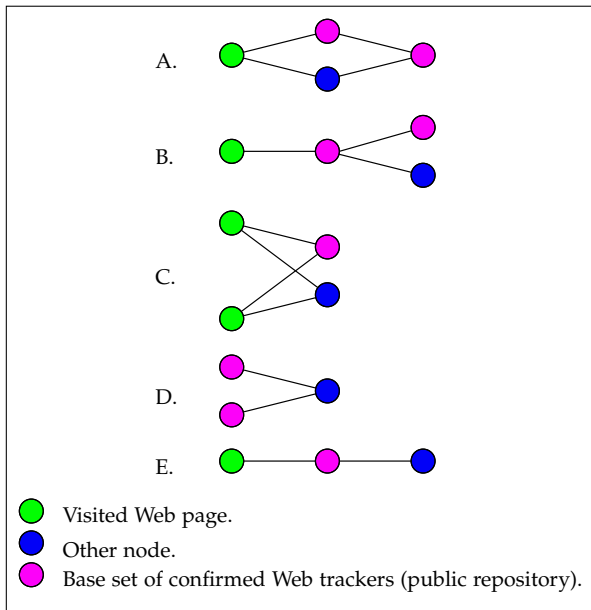


Figure 3.3: Graph mining rules for detecting web tracking in a graph database, e. g., Neo4j. The Cypher Query Language (Cypher) can express these patterns effectively [Van Eijk, 2011b, p. 41].

Van Eijk [2011b] presented nine findings (not reproduced here). On the one hand, his analysis showed two distinct indicators for web tracking:

- (1) the number of filtered nodes as a percentage of the total number of nodes, and
- (2) a cluster coefficient describing the interconnectedness between nodes expressed in the number of links per node.

On the other hand, I formulated five distinct subgraph patterns representing an indication for web tracking. In Figure 3.3 the subgraph patterns are denoted by A, B, C, D, and E.

Van Eijk [2011b] showed that the graph-mining rules represented (1) frequent patterns of web tracking and (2) enabled the detection of missing values through small data. Three colors were used to differentiate between:

- (1) the media website visited (green nodes),
- (2) a web tracking domain not found in a public repository of confirmed web trackers (blue nodes), and
- (3) web-tracking domains found in a public repository of confirmed web trackers (purple nodes).

A: ILLUSTRATION AND FORMALIZATION OF SMALL DATA

To illustrate the concept of *small data* we draw an analogy between DNA and web tracking. Just as the essence of a string of DNA is situated in its small data - i. e., a sequence of nucleotides shaping the right-handed double helix - the essence of the information in the HTTP header fields is situated in its small data.

All subgraph patterns are examples of web-tracking phenomena derived from data analysis from a *host* perspective. The patterns enable the identification of hosts that are unknown for their ability to understand an end-user's browsing interests to target ads and personalize web services accordingly.

Below we formalize each pattern shown in Figure 3.3. Let us start formalizing a subgraph by adopting the definition of a subgraph as proposed by [Riesen et al., 2010] as our Definition 3.5.

DEFINITION 3.5: „SUBGRAPH. Let $g_1 = (V_1, E_1, \mu_1, \nu_1)$ and $g_2 = (V_2, E_2, \mu_2, \nu_2)$ be graphs. Graph g_1 is a subgraph of g_2 , denoted by $g_1 \subseteq g_2$, if

- (1) $V_1 \subseteq V_2$,
- (2) $E_1 \subseteq E_2$,
- (3) $\mu_1(u) = \mu_2(u)$ for all $u \in V_1$, and
- (4) $\nu_1(e) = \nu_2(e)$ for all $e \in E_1$.” [Riesen et al., 2010, p. 220]

The definition of a subgraph enables us to formalize each pattern that is shown in Figure 3.3. The corresponding formalizations of Subgraph A, Subgraph B, Subgraph C, Subgraph D, and Subgraph E are given by Definition 3.6 to Definition 3.10.

DEFINITION 3.6: SUBGRAPH A is defined to be a subgraph denoted by $g = (V, E, \mu, \nu)$ with four nodes (V_A, V_B, V_C, V_D) and four edges ($E_1: V_A \times V_B, E_2: V_B \times V_D, E_3: V_A \times V_C, E_4: V_C \times V_D$), and three distinct node properties ($\mu_A=\text{green}, \mu_B=\text{purple}, \mu_C=\text{blue}, \mu_D=\text{purple}$).

The host visited (V_A) contains two confirmed tracker nodes (V_B, V_D) and one other node (V_C). We consider V_C a tracker node when the interconnectedness between V_A, V_C , and V_D is similar to the interconnectedness between V_A, V_B , and V_D .

DEFINITION 3.7: SUBGRAPH B is defined to be a subgraph denoted by $g = (V, E, \mu, \nu)$ with four nodes (V_A, V_B, V_C, V_D) and three edges ($E_1: V_A \times V_B, E_2: V_B \times V_C, E_3: V_B \times V_D$), and three distinct node properties ($\mu_A=\text{green}, \mu_B=\text{purple}, \mu_C=\text{purple}, \mu_D=\text{blue}$).

The host visited (V_A) contains two confirmed tracker nodes (V_B, V_C) and one other node (V_D). We consider V_D a tracker node when the interconnectedness between V_A, V_B , and V_D is similar to the interconnectedness between V_A, V_B , and V_C .

DEFINITION 3.8: SUBGRAPH C is defined to be a subgraph denoted by $g = (V, E, \mu, \nu)$ with four nodes (V_A, V_B, V_C, V_D) and four edges ($E_1: V_A \times V_C, E_2: V_A \times V_D, E_3: V_B \times V_C, E_4: V_B \times V_D$), and three distinct node properties ($\mu_A=\text{green}, \mu_B=\text{green}, \mu_C=\text{purple}, \mu_D=\text{blue}$).

The hosts visited (V_A and V_B) contain a confirmed tracker node (V_C) and one other node (V_D). We consider V_D a tracker node when the interconnectedness between V_A, V_D , and V_B is similar to the interconnectedness between V_A, V_C , and V_B .

DEFINITION 3.9: SUBGRAPH D is defined to be a subgraph denoted by $g = (V, E, \mu, \nu)$ with three nodes (V_A, V_B, V_C) and two edges ($E_1: V_A \times V_C, E_2: V_B \times V_C$), and two distinct node properties ($\mu_A=\text{purple}, \mu_B=\text{purple}, \mu_C=\text{blue}$).

If the other node (V_C) is connected to two confirmed tracker nodes (V_A and V_B) then V_C can be considered a tracker node as well.

DEFINITION 3.10: SUBGRAPH E is defined to be a subgraph denoted by $g = (V, E, \mu, \nu)$ with three nodes (V_A, V_B, V_C) and two edges ($E_1: V_A \times V_B, E_2: V_B \times V_C$), and three distinct node properties ($\mu_A=\text{green}, \mu_B=\text{purple}, \mu_C=\text{blue}, \mu_D=\text{purple}$).

If the other node (V_C) is interconnected to the visited website (V_A) via a confirmed tracker node (V_B) then V_C can be considered to be a tracker node as well.

B: INDUCED SUBGRAPHS AND ITS APPLICATIONS

We continue for a while our reflection on the findings by Van Eijk [2011b]. Riesen et al. [2010] proposed a definition of an induced subgraph by replacing the second condition in Definition 3.5 as follows.

DEFINITION 3.11: „INDUCED SUBGRAPH. Let $g_1 = (V_1, E_1, \mu_1, \nu_1)$ and $g_2 = (V_2, E_2, \mu_2, \nu_2)$ be graphs. Graph g_1 is a subgraph of g_2 , denoted by $g_1 \subseteq g_2$, if

- (1) $V_1 \subseteq V_2$,
- (2) $E_1 = E_2 \cap V_1 \times V_1$,
- (3) $\mu_1(u) = \mu_2(u)$ for all $u \in V_1$, and
- (4) $\nu_1(e) = \nu_2(e)$ for all $e \in E_1$.” [Riesen et al., 2010, p. 220].

We remark that the concept of an induced subgraph can be applied to our patterns. From Definition 3.11 we can derive two new subgraph concepts. First, it follows that subgraph E is an induced subgraph of graph B. Subgraph E can be obtained from graph B by removing a node.

Second, it follows that subgraph D is an induced subgraph of graph A. This is not obvious at first sight. It is necessary first to determine a blue colored (and therefore uncodified) node as a tracker node, then it becomes clear that subgraph D can be obtained from graph A by removing a node.

C: MOTIF ENUMERATION AND COUNTING ALGORITHMS

Here, we remark that the five subgraphs above are defined manually (Definition 3.6 – Definition 3.10). However, recent developments in network science show that such patterns can be traced

automatically by the application of motif enumeration and counting algorithms (see, e. g., Benson, Gleich, and Leskovec [2016]).²¹⁵

D: CONCLUSIONS ON THE FORMALIZATION OF GRAPH-MINING RULES

In conclusion, our reflection on the formalization of the findings by Van Eijk [2011b] lead us to the following five advantages of adding color to the graph analysis.

- (1) The graph mining rules enable the visualization of web tracking in a force-directed graph.
- (2) The patterns provide a means of detecting missing values, i. e., unknown web trackers.
- (3) The patterns can be formalized.
- (4) The patterns are early examples of subgraph isomorphism applied to the problem of web tracking.²¹⁶
- (5) The patterns enable an understandable graph-based analysis. The rules are queries that can be expressed in a query language such as Neo4j's Cypher Query Language.²¹⁷

3.3.3 *Extended web-tracking model*

Subsequently, Van Eijk [2012] improved the generic web-tracking model. Van Eijk [2011b, p. 32] had already established that a typical (unique) tracking identifier for RTB was at least a ten-digit variable. For an improved model (henceforth: extended web-tracking model) we add four new elements, viz.

- (1) filtering the use of high-entropy identifiers in *any* HTTP header fields, e. g., the ETag header field,
- (2) adding high-entropy as a property to node C (Figure 3.2),
- (3) directing the graph, and
- (4) adding meta data as properties to the edges (e. g., which website sets a UID in a HTTP cookie, or which websites received the UID).

An example of the extended web-tracking model is given in Figure 3.4 (next page). I presented and discussed the example as

²¹⁵ I am indebted to Takes [2018] for this insight.

²¹⁶ Infra Definition 3.17.

²¹⁷ See, e. g., Robinson et al. [2015]; N. White [2014].

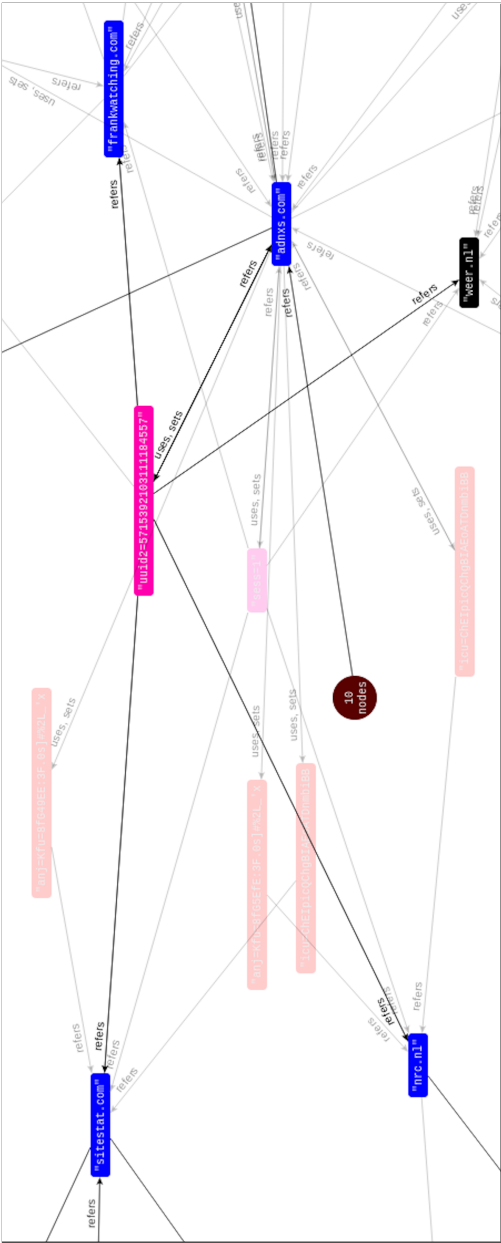


Figure 3.4: The RTB cookie named 'uuid2' with value '5715392183111164557' is set by the domain 'adnxs.com'. An RTB cookie with the same value was connected to the four domains we visited (1) 'nrc.nl', (2) 'weer.nl', (3) 'sitestat.com', and (4) 'nrc.nl' [Van Eijk, 2012].

an early empirical result of web tracking with UID-cookies at the International Association of Privacy Professionals (IAPP) Europe Data Protection Congress 2012. Figure 3.4 shows a directed property graph.

The graph is compiled with the user friendly Neo4j browser. It enables us to easily query and visualize our research data (Crawl1, Table 3.2). Figure 3.4 contains six nodes in focus ('sitetat.com', 'nrc.nl', 'uuid2', 'frankwatching.com', 'adnxs.com', and 'weer.nl'), one conflated node labeled '10 nodes', and five other (less important) nodes. The Neo4j-browser interface enables us to zoom in on conflated nodes by a mouse click.

At the right of the graph we see domain 'adnxs.com' (AppNexus) as a node in blue color. Although AppNexus is referred to by more than 43 nodes, the focus in our visualization is on the edge between the nodes labeled 'adnxs.com' and 'uuid2=5715392103111184557'. The properties of the directed edge are:

- (A) 'uses',
- (B) 'sets', and
- (C) 'refers'.

The property 'uses' means reading the UID-cookie, whereas the 'sets' stand for setting the UID-cookie in the browser. The property 'refers' denotes a reference to AppNexus in the HTTP header. The central focus on the UID-cookie allows us to see the interconnection by the tracking cookie on the following four domains (blue nodes): (1) 'nrc.nl', (2) 'weer.nl', (3) 'frankwatching.com', and (4) 'sitetat.com'.

3.3.4 High-entropy identifiers

The prevailing question now is: what type of tracking identification do we use for our experiments. Van Eijk [2012] used entropy ([Shannon, 1948]) as a filter for data reduction. The filter uses high-entropy as a criterion for web tracking. As such, it is an example that illustrates our way of modeling metadata.

The main reason to base the experiments on high-entropy cookie IDs was rooted on the idea proposed by Eckersley [2010a], namely that high browser entropy increases the risk identification by a (unique) browser fingerprint (see Section 2.4). Since cookies played (and still do today) a central role in web tracking, the aim of the experiments was to differentiate between tracking cookies

and non-tracking cookies with entropy as a measure. The Shannon entropy is defined as follows.

DEFINITION 3.12: SHANNON ENTROPY is defined to be the Shannon entropy H of a random variable X , with a logarithmic base value j of 2, with outcomes $\{x_1, \dots, x_n\}$, and a probability $P(x_i)$.

The Shannon entropy is given by:

$$H(X) = - \sum_{i=1}^n P(x_i) \log_j P(x_i) \quad (3.1)$$

We define the term high-entropy identifier as follows.

DEFINITION 3.13: HIGH-ENTROPY IDENTIFIER. A Unique Identifier (UID) containing at least 13 bits of (Shannon) entropy, which is sufficient to identify an individual in a group of 5,000 people.

Below we provide (A) three examples of using entropy, (B) an algorithmic description, and (C) observations.

A: THREE EXAMPLES OF USING ENTROPY

The size of the group is consistent with the compliance framework for Do Not Track (DNT) (henceforth: EFF DNT policy) of the Electronic Frontier Foundation (EFF).²¹⁸ The EFF DNT policy contains a normative privacy framework that aims to limit the reconstruction of reading habits or online activity of people using the web. Under the EFF DNT policy, data groups of fewer than 5,000 individuals or devices are considered as identifiable end-users rather than anonymous data. Below we discuss three examples: (A1) a perfect die, (A2) a language code, and (A3) identifiers with multiple elements.

A1: FIRST EXAMPLE, A PERFECT DIE

The Shannon entropy of a single roll with a perfect die is determined by its six sides. Because the chance $P(x_i)$ of the perfect die rolling each number is equal ($\frac{1}{6}$), the entropy H is ~ 2.58 bits of

²¹⁸ See EFF [2015].

binary information to encode the outcome of a single roll with the die.

A2: SECOND EXAMPLE, A LANGUAGE CODE

A second example to illustrate the working of our tracking identification reads as follows. It is language oriented and thus more related to web tracking. Web developers may use a language code in a cookie to store a language preference. ISO [2013, ISO 639-1] defines abbreviations for languages, e. g., 'nl' for the Netherlands. The entropy of language codes is mathematically determined by the number of common languages. The Firefox add-on Privacy-Badger relies on a list of 100 common language codes.²¹⁹ Therefore, the entropy of the listed language codes is ~ 6.64 bits.

A3: THIRD EXAMPLE, IDENTIFIERS WITH MULTIPLE ELEMENTS

A third example sees to identifiers consisting of multiple elements. Only when the probability distributions of each element are independent, the entropy values can be totaled. To understand the dependency of elements, let us consider an example with three elements, e. g., (1) the browser's *local time*, (2) the browser's *time zone*, and (3) the browser's *geolocation* from where the end-user is browsing.

The question here is: can the entropy values of these three elements be totaled? In case the two browser properties *local time* and *time zone* are kept to its default localized values and the *location* where the end-user is browsing from is known through a geolocation lookup of the IP address, the probability distributions of the three browser properties are not independent. The reason for this is that when one knows the *local time* and the *time zone*, knowing the *geolocation* does not increase the entropy.²²⁰

B: AN ALGORITHMIC DESCRIPTION

Van Eijk [2012] applied Shannon's entropy to filter high-entropy identifiers. A threshold was used to differentiate between functional identifiers and tracking identifiers with a high-entropy. Preliminary results of empirical exploratory research were presented

²¹⁹ Cf. Eckersley et al. [2014]. See also Eckersley [2010b; 2015].

²²⁰ Cf. Eckersley [2010b]. See also Koot, Mandjes, Van't Noordende, and De Laat [2011] and Koot [2012].

at the 2012 IAPP Europe Data Protection Congress [Wefers Beting, Van Eijk, & Wagner, 2012].

Algorithm 3.1 [Van Eijk & Terra, 2012] describes the implementation of Shannon’s entropy.

Algorithm 3.1: Calculating the high-entropy value of a string s .

input : A string s containing an identifier
output: High-entropy value

```

1  $H \leftarrow 0$ ;
2 forall unique characters  $c$  in  $s$  do
3    $P_c \leftarrow s.\text{count}(c)/\text{len}(s)$ ;
4    $H \leftarrow H + P_c \log_2 P_c$ ;
5 return  $-\text{len}(s)H$ ;
```

The high-entropy value of a random string s was calculated as follows. For a string s , each character c is treated as an independent variable. The probability $P(x_i)$ of finding a certain character is given by the number of occurrences of that character in the string divided by the length of the string. The high-entropy value is calculated by multiplying the entropy H with the length of the string. Van Eijk and Terra [2012] implemented the concept in Python.

Determining whether an identifier qualifies as a tracking cookie is a challenge.²²¹ The main issue in a heuristic approach is striking the right balance between (1) false positives (see Definition 3.14) and (2) false negatives (see Definition 3.15).

DEFINITION 3.14: A FALSE POSITIVE is an identifier that qualifies as an heuristically determined high-entropy identifier, but in reality it is not used for web tracking.

For instance, we mention a HTTP cookie containing the version number of a JavaScript file. Reducing the number of false positives requires improving the heuristic approach.

²²¹ We refer to, e.g., Adams et al. [2012], Eckersley et al. [2014], and Reisman, Englehardt, Eubank, Zimmerman, and Narayanan [2014].

DEFINITION 3.15: A **FALSE NEGATIVE** is an identifier that was missed by an heuristic approach. False negatives point to a lack of granularity in the heuristic model.

C: OBSERVATIONS

Asghari, Van Eijk, Englehardt, Narayanan, and Winter [2016] applied Algorithm 3.1 within the framework of their EU-cookie transparency & accountability project (CookieTAP). One of the questions we asked ourselves was: how does entropy as a metric compare to counting third-party cookies? The assumption here was that third-parties do store, e.g., (1) IDs for the end-user and (2) event tracking during session in HTTP cookies.

To answer the question, we set up stateless shallow crawls with OpenWPM [Englehardt et al., 2014].²²² The results of our analysis are shown in Figure 3.5. Their meaning is: it is almost linear (say $\alpha = 40^\circ$ in $y = \alpha x$). This means no difference has been found between the two measures.

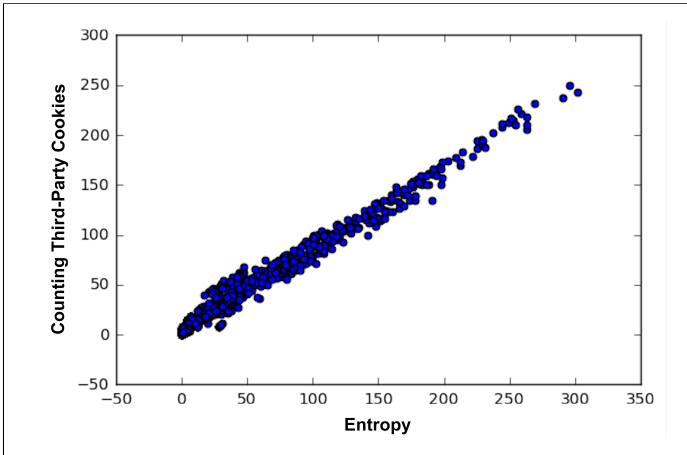


Figure 3.5: Entropy versus Counting Third-Party Cookies (CTPC):
Spearman's correlation coefficient = 0.95758713547808916,
pvalue = 0.0 [Asghari et al., 2016].

²²² We set up stateless shallow crawls with OpenWPM [Englehardt et al., 2014] (name: crawl_20161201.1906_alex10x25_v9). We retained a copy of the crawl data and the Jupyter [Perez, 2001] notebooks in a (private) github repository. URL: <https://github.com/hadiasghari/EUcookies> (2 December 2017).

For closer inspection, we visited a small sample of 140 websites from nine countries, i. e., (1) Austria, (2) Germany, (3) France, (4) United Kingdom, (5) United States of America, (6) Italy, (7) the Netherlands, (8) Poland, and (9) Romania.²²³ To get access from inside a EU-country, we crawled with OpenVPN [Yonan, 2001] endpoints from the IP-address of the Center for Information Technology Policy at Princeton University.²²⁴ Surprisingly, also here we found no difference between the two metrics. The Spearman's correlation coefficient of our dataset is close to one.²²⁵

Van Eijk and Cortesi [2016] also re-investigated the issue and confirmed the result. Moreover, they found:

- (1) entropy (Algorithm 3.1) performed poorly on short cookie values,
- (2) entropy did not match intuition for structured data in longer cookie values, and
- (3) entropy did not distinguish session cookies that did not persist in the browser.

3.3.5 *A small-data model*

After the negative result above, we continued our aim to further improve the extended web-tracking model. Therefore, we propose to formalize web tracking as an exact graph matching problem. This means that we are interested in subgraph isomorphism. The course of this subsection is as follows: (A) subgraph isomorphism, (B) customization, and (C) Ghostery web-tracking classification example.

²²³ The 140 websites were picked from the top-Alexa domain ranking: '.at' (8), '.bz' (1), '.co.uk' (6), '.de' (10), '.fr' (9), '.gov.uk' (1), '.ie' (7), '.it' (11), '.net' (5), '.nl' (7), '.org' (4), '.pl' (16), '.ro' (9), '.ru' (3), '.se' (8), and '.com' (35).

²²⁴ The OpenVPN endpoints were provided by the companies CyberGhost, Privax (HideMyAss), and FoxyProxy.

²²⁵ See also Turcios Rodríguez [2018, pp. 88–89] who re-investigated the issue and measured the correlation between unique cookies names and unique third-party domains ($n = 35,000$ websites, Spearman's correlation coefficient = 0.88792705220847501, $pvalue = 0.0$). Moreover, she measured the correlation between the use of JavaScript and unique third-party domains ($n = 35,000$ websites, Spearman's correlation coefficient = 0.91435850137104757, $pvalue = 0.0$).

A: SUBGRAPH ISOMORPHISM

Below we adopt the definition of graph isomorphism (Definition 3.16) and subgraph isomorphism (Definition 3.17) as proposed by Riesen et al. [2010]. The definitions are as follows.

DEFINITION 3.16: „GRAPH ISOMORPHISM. Let us consider two graphs denoted by $g_1 = (V_1, E_1, \mu_1, \nu_1)$ and $g_2 = (V_2, E_2, \mu_2, \nu_2)$. A graph isomorphism is a *bijective* function $f : V_1 \rightarrow V_2$ satisfying

- (1) $\mu_1(u) = \mu_2(f(u))$ for all nodes $u \in V_1$,
- (2) for each edge $e_1 = (u, v) \in E_1$, there exists an edge $e_2 = (f(u), f(v)) \in E_2$ such that $\nu_1(e_1) = \nu_2(e_2)$, and
- (3) for each edge $e_2 = (u, v) \in E_2$, there exists an edge $e_1 = (f^{-1}(u), f^{-1}(v)) \in E_1$,

such that $\mu_1(e_1) = \nu_1(e_1)$. Two graphs are called isomorphic if there exists an isomorphism between them.” [Riesen et al., 2010, pp. 221–222] (emphasis added)

DEFINITION 3.17: „SUBGRAPH ISOMORPHISM. Let $g_1 = (V_1, E_1, \mu_1, \nu_1)$ and $g_2 = (V_2, E_2, \mu_2, \nu_2)$ be graphs. An *injective* function $f : V_1 \rightarrow V_2$ from g_1 to g_2 is a subgraph isomorphism if there exists a subgraph $g \subseteq g_2$ such that f is a graph isomorphism between g_1 and g .” [Riesen et al., 2010, p. 223] (emphasis added)

Bijection is a mathematical term that describes a one-to-one correspondence, or the exact matching of two graphs. The term injection describes a one-to-one function that preserves distinctness, or exact matching of a subgraph.

B: CUSTOMIZATION

Customized graph data models (henceforth: small-data models) are particularly suitable when it comes to solving graph matching problems. The key objective for modeling the small data is to include all key properties needed for subgraph isomorphism. The properties correspond with specific properties (metadata) of the research data. Small-data modeling by feature extraction from the HTTP headers enables encoding of research questions in a

way that feels like natural speech. The metadata in the data flow allow us to model the small data. We provide a telling example: (B1) an example of customization, including (B2) an algorithmic background of the example.

B1: AN EXAMPLE OF CUSTOMIZATION

To illustrate small-data modeling, we turn to event-level tracking for email. The example stems from a personal account. Recently I signed up for a trial period for cloud-based transcription software.²²⁶ At the end of the trial period I received a notification email containing a friendly reminder to purchase a software license.²²⁷ The email reminder contained a hyperlink with a HTTP redirect.²²⁸ Instead of taking me straight to the login page for my account, the login page loaded after having been redirected from a URI recording the click event.

URI redirection is a common HTTP mechanism to instruct the browser to load a resource that is located somewhere else (cf. Fielding and Reschke [2014b, RFC 7231, Section 6.4]). The HTTP response status code 302 indicates URI redirection. HTTP redirects can also be used to uniquely identify end-users, a tracking technique known as ‘301 moved permanent redirect tracking’.²²⁹

A small-data model for event-level tracking for email is shown in Figure 3.6. The aim of the small-data model is to investigate, e. g., the following question: which URIs requested a resource from a third party and were redirected to a resource from another third-party?

The small-data model illustrates how the phenomenon of URI redirection could be encoded in a directed graph model. The research data stems from the HTTP headers.

The small-data model (Figure 3.6) has edge labels and node labels. The labels ‘REQUEST’ (encoding of the HTTP GET request) ,

226 URL: <http://transcribe.wreally.com/login> (26 December 2015).

227 „You can do so [purchasing a license] by logging into your Transcribe account and visiting your account page”.

228 The URI was as follows URL: [http://emails.transcribe.wreally.com/track/click/30097249/transcribe.wreally.com?p=eyJzIjo1VDIwc1J\(...\).SJ9](http://emails.transcribe.wreally.com/track/click/30097249/transcribe.wreally.com?p=eyJzIjo1VDIwc1J(...).SJ9) (26 December 2015) (shortened).

229 Mozilla [2018b]: „301 moved permanent redirect tracking: if you visit a site, it might load a resource that has a 301 redirect. The resource can redirect you to URL that is unique to you. Then, the next time you see the original resource, your browser will load the unique URL from the cache, and fetch the resource from there. Making you uniquely identified.”

'REDIRECT' (encoding of the HTTP 302 redirect), and 'NEXT' (encoding of two consecutive URI visits) are edge properties. Above we have modeled HTTP redirection in a graph-based object. It enables us to formalize the subgraph with the following definition.

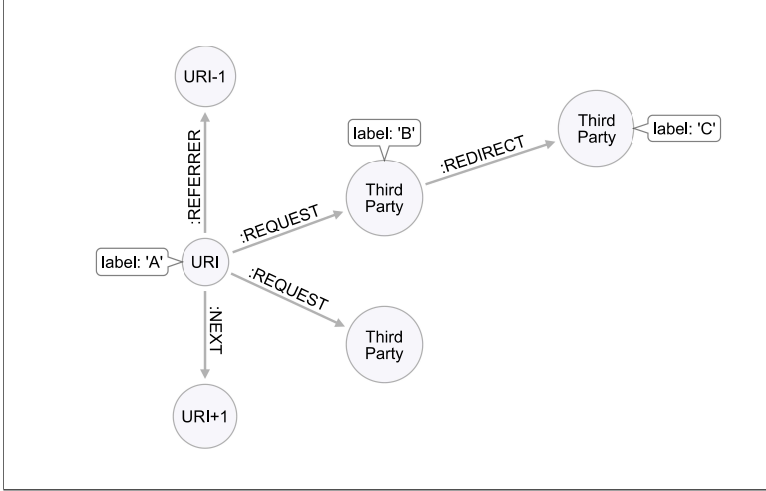


Figure 3.6: Example of a small-data model to investigate HTTP redirection through subgraph isomorphism.

DEFINITION 3.18: HTTP REDIRECTION. Let us consider a directed subgraph denoted by $g = (V, E, \mu, \nu)$ with six nodes ($V_1, V_2, V_3, V_4, V_5, V_6$), five edges ($E_1: V_1 \leftarrow V_2, E_2: V_2 \rightarrow V_3, E_3: V_2 \rightarrow V_4, E_4: V_2 \rightarrow V_5, E_5: V_4 \rightarrow V_6$), six distinct node properties ($\mu_1 = \text{URI-1}, \mu_2 = \text{URI}, \mu_3 = \text{URI+1}, \mu_4 = \text{'Third Party'}, \mu_5 = \text{'Third Party'}, \mu_6 = \text{'Third Party'}$), and five distinct edge properties ($\nu_1 = \text{'REFERRER'}, \nu_2 = \text{'NEXT'}, \nu_3 = \text{'REQUEST'}, \nu_4 = \text{'REQUEST'}, \nu_5 = \text{'REDIRECT'}$)

HTTP redirection is defined to be an induced subgraph g' of g , if g' consists of the three nodes (V_2, V_4, V_6) and the two edges (E_3, E_5).

B2: ALGORITHMIC BACKGROUND OF THE CUSTOMIZATION EXAMPLE

To better understand the practicality of encoding questions in a small-data model we look at the place where to look in the research data. We restrict the example (Listing 3.3) to URI redirection. The induced subgraph we are interested in consists of the nodes labeled A, B, and C in Figure 3.6. The (simplified) HTTP request header is given in Listing 3.3.

Listing 3.3: HTTP request header with a GET request to a web resource.

```

1 GET /index.html HTTP/1.1
2 Host: www.domain-a.nl
3 ...

```

In response to the HTTP request above, the server responds with a HTTP 302 status code. The response is given below in Listing 3.4.

Listing 3.4: HTTP response header containing a HTTP 302 status code and the HTTP Location header field containing the redirect URI.

```

1 HTTP/1.1 302 Moved Temporarily
2 Location: http://www.third-party.nl
3 ...

```

The HTTP status code, and the domain names of the URIs - 'uri-1.nl' (Listing 3.3, r. 2) and 'third-party.nl' (Listing 3.4, r. 2) - are node properties of their respective nodes. A Cypher query encoding the example question could be based on the isomorphic pattern, given in Listing 3.5.

Listing 3.5: Subgraph isomorphism: Matching pattern for a Cypher query.

```

1 (URI) -[REQUEST]->(Third Party) -[REDIRECT]->(Third Party)

```

Of course, small-data models for WPM research do not have to be restricted to HTTP-header information. In summary, what we intended to show is the possibility that metadata in the data flow

allows us to model the small data. We believe that this point has been made.

C: GHOSTERY WEB-TRACKING CLASSIFICATION EXAMPLE

A topic that is somewhat related with the small-data model is Ghostery [Pierce, 2010]. Below we show a tracking protection technology based on the generic web-tracking model.²³⁰

A relevant development for our research is the possibility to block web traffic. In the example below we demonstrate the limitations of blocking web tracking when using the well-known browser extension Ghostery [Pierce, 2010]. Ghostery is (1) a transparency tool for web tracking, but (2) the extension gives end-users (limited) control over blocking tracking elements in a webpage. So, Ghostery blocks web-tracking elements with a blacklist approach. The effectiveness of blacklist blocking has been studied by WPM scholars, see, e. g., Mayer and Mitchell [2012, p. 12].

Below we briefly discuss (C1) the modern tracking-protection technologies, and (C2) the generic web-tracking model with node classification.

C1: TRACKING PROTECTION TECHNOLOGY

Nowadays browser manufacturers endow their browsers with built-in tracking protection. Of course, end-users are also allowed to configure the browser with (experimental) tracking extensions. Although not enabled by default, shipping a browser with tracking protection technology is a major step in terms of end-user empowerment, since then the underlying privacy question (Section 1.2) is at stake. We can subdivide the underlying technology into three types:

- (1*) host based,
- (2*) regular expressions, and
- (3*) tracking classification.

Below we study this protection technology in some details. For a proper understanding, we provide brief descriptions for each technology.

²³⁰ Generic web-tracking model (Figure 3.2). For the examples I drew inspiration from N. White [2014] who demonstrated the use of RNeo4j.

1*: **HOST BASED** Microsoft has been a proponent of host-based tracking-protection lists. Although Microsoft does not maintain a list itself, end-users can configure the Internet Explorer with host based lists available online.²³¹ Mozilla ships Firefox with a tracking protection list provided by its partner Disconnect [Kennish et al., 2013].²³²

2*: **REGULAR EXPRESSION** is a programming method for locating specific character strings embedded in character text [Thompson, 1968]. A well-known list based on regular expressions is Easylist [Petnel, 2005]. *Ghostery* also uses this tracking protection technology.

3*: **TRACKING CLASSIFICATION** Apple's Intelligent Tracking Prevention (ITP) is a state of the art technology based on machine learning. Apple ships its Safari browser with ITP (iOS 11). Apple's algorithmic approach is based on three indicators for tracking (cf. Wilander [2017]).²³³

- (1) a subresource on the number of unique domains, e. g., a tag (see Subsection 4.3.8),
- (2) a subframe on the number of unique domains, e. g., a Google SafeFrame, and
- (3) the number of unique domains redirected to (see Figure 3.6).²³⁴

C2: GENERIC WEB-TRACKING MODEL WITH NODE CLASSIFICATION

In Figure 3.7 and Figure 3.8 (p. 126) we demonstrate the practicality of small-data modeling with node classification.

231 URL: [https://msdn.microsoft.com/en-us/library/hh273399\(v=vs.85\).aspx](https://msdn.microsoft.com/en-us/library/hh273399(v=vs.85).aspx) (27 November 2017).

232 See, URL: <https://support.mozilla.org/en-US/kb/tracking-protection> (27 November 2017). The latest version of the blacklist is available online. URL: <https://github.com/disconnectme/disconnect-tracking-protection/blob/master/services.json> (27 November 2017).

233 The original quotation is as follows: „A machine learning model is used to classify which top privately-controlled domains have the ability to track the user cross-site, based on the collected statistics. Out of the various statistics collected, three vectors turned out to have strong signal for classification based on current tracking practices: [1] subresource under number of unique domains, [2] sub frame under

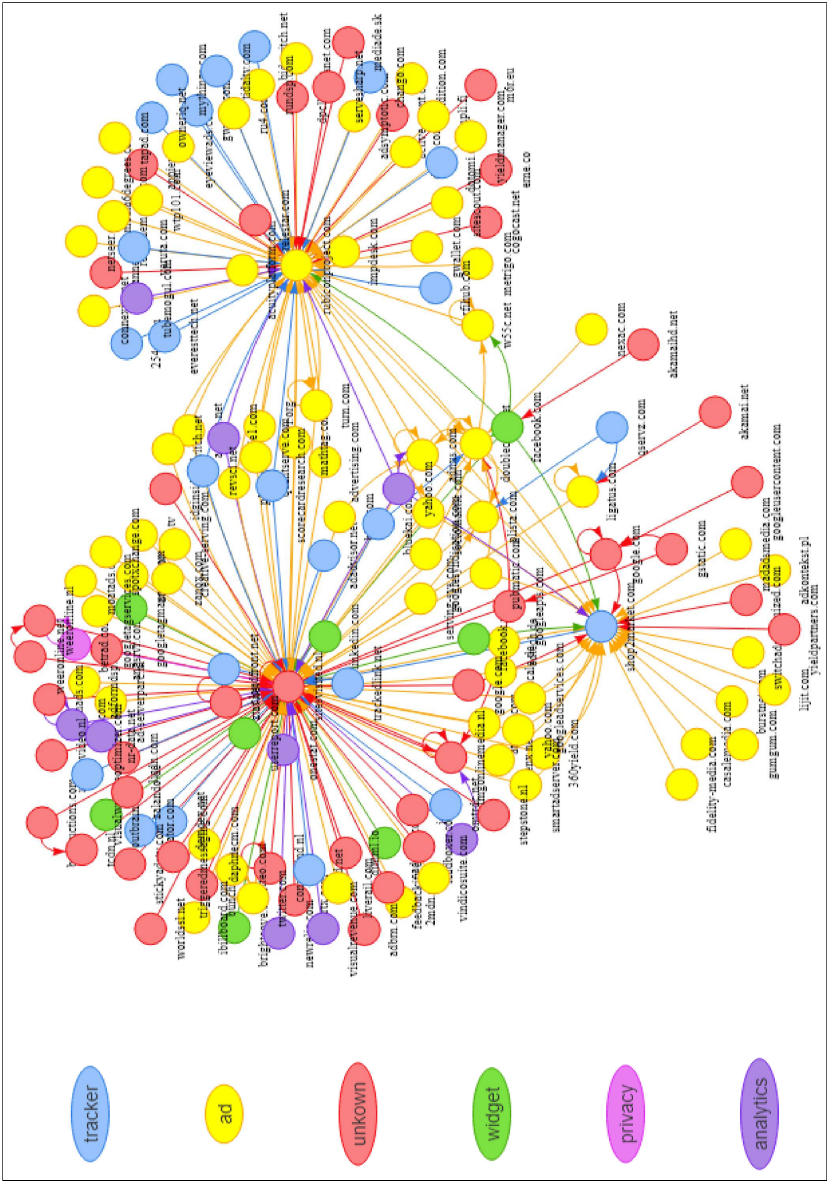


Figure 3.7: Generic web-tracking model with Ghostery-node classification (N=158 nodes and E=228 edges).

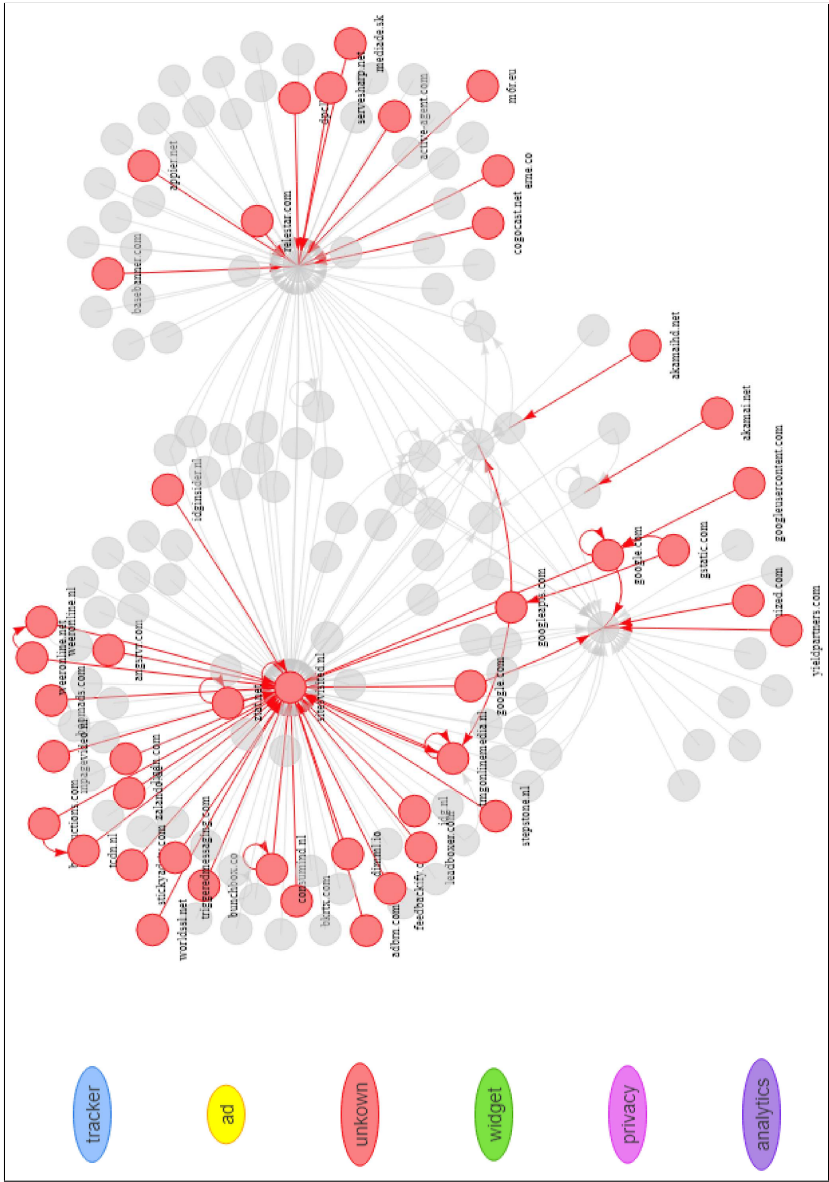


Figure 3.8: Ghostery-node classification: missing values.

We analyzed Ghostery's tracking classification. The Ghostery browser extension contains metadata to classify web tracking.²³⁵ The result of our analysis is visualized in an undirected property graph. Figure 3.7 shows all categories whereas Figure 3.8 highlights the missing values (nodes in red color denoted by the category 'unknown').

The graph represents data from manually visiting four Dutch websites.²³⁶ We applied the generic web-tracking model to the data with the R-package *visNetwork* [Thieurmél, 2016].²³⁷

To make the graph more meaningful, we combined the individual nodes of the four sites visited into one node ('sitesvisited.nl').²³⁸ The graph consists of 158 nodes. The Ghostery metadata enables us to add a color as a property for the tracking category to the nodes.

The six categories denoted by Ghostery's metadata are:

- (1) tracker (24 nodes),
- (2) ad (72 nodes),
- (3) unknown (44 nodes),²³⁹
- (4) widget (8 nodes),²⁴⁰
- (5) privacy (1 node),
- (6) analytics (9 nodes).

In summary, Ghostery is known for providing transparency in web tracking and giving its end-users more control over their (private) data when they browse the web. Our experiment shows that 44 nodes (27 %) still received (personal) data because they were

number of unique domains, and [3] number of unique domains redirected to." [Wilander, 2017]

234 I am indebted to O'Neill [2017] who confirmed Apple's algorithmic approach to Apple's ITP.

235 The metadata was released by its owner Cliqz under an open source license in March 2018. URL: <https://github.com/ghostery/ghostery-extension/tree/master/databases> (9 March 2018).

236 The websites visited are: (1) webwereld.nl, (2) www.debijenkorf.nl, (3) www.zalando.nl, and (4) www.telegraaf.nl. We retained the data with Mitmproxy (filename: 150324-bijenkorf-webwereld-test.mitm).

237 We analyzed the data in R [R Core Team, 2014]. We shortened the URL to the domain name and added it as a property to the nodes. We used the referrer field (HTTP request header) and location field (HTTP response header) as information for the edges. The *visNetwork* [Thieurmél, 2016] package comes with a feature to create a web application which simplifies further analysis of the graph.

238 Infra our definition for graph refactoring (see Definition 4.13).

239 We colored the central node 'sitesvisited.nl' red.

240 For example a Facebook-like button

classified as missing values ('unknown') by the browser extension. In general, the classification problem is an illustration of the shortcomings of a blacklist approach. We mark that a blacklist approach to prevent web tracking will not be 100% complete.²⁴¹ We refer to, e.g., Englehardt et al. [2018, pp. 9–10], Merzdovnik, Huber, Buhov, Nikiforakis, Neuner, Schmiedecker, and Weippl [2017], Traverso, Trevisan, Giannantoni, Mellia, and Metwalley [2017], and Mazel et al. [2017] (who proposed a reliable methodology for privacy protection comparison, including blacklist approaches, and extensively compared a wide set of privacy protection techniques).

We remark that Ghostery acknowledged the gap in tracking protection (see Konrad [2017]).²⁴² The company improved their product's tracking-protection capabilities with an algorithmic approach to rewriting UIDs as a measure to reduce the privacy risk of tracking end-user behavior.²⁴³

To better understand the missing values, we will further investigate the RTB landscape with its multiple actors and their interrelationships in Chapter 4. There we show two intricacies which we handel with two M-steps (see Subsection 4.4.3). Finally, we remark that Ghostery acknowledged the gap in tracking protection [Konrad, 2017].²⁴⁴ The company improved their product's tracking-protection capabilities with an algorithmic approach to rewriting UIDs as a measure to reduce the privacy risk of tracking end-user behavior.²⁴⁵

²⁴¹ The same applies to, e.g., Mozilla's (always on) blacklist approach for tracking protection [Mozilla, 2018a].

²⁴² The full quotation is as follows: „But as tracking companies constantly update existing trackers and add new ones, it would be impossible to detect and block all trackers on the internet.”

²⁴³ The technology was developed by Cliqz: „Ghostery 8's enhanced privacy protection detects and overwrites any uniquely identifying data points being passed to a third party - even trackers that aren't caught by Ghostery's comprehensive blacklist” [Konrad, 2017].

²⁴⁴ The full quotation is as follows: „But as tracking companies constantly update existing trackers and add new ones, it would be impossible to detect and block all trackers on the internet.”

²⁴⁵ The technology was developed by Cliqz: „Ghostery 8's enhanced privacy protection detects and overwrites any uniquely identifying data points being passed to a third party - even trackers that aren't caught by Ghostery's comprehensive blacklist” [Konrad, 2017].

3.4 CHAPTER CONCLUSIONS

In this chapter we presented our GBMA by investigating three parts: data collection, data reduction, and data modeling.

In Part 1, data collection, we proposed a new 3-step procedure (C-steps) of data flow collection (see Subsection 3.1.3). In Part 2, data reduction, we again proposed a 3-step procedure (R-steps) of reducing the research data (see Section 3.2). In Part 3, data modeling, we proposed a small-data model, in which the metadata are stored (see Section 3.3).

Here, we remark that an analysis of the strengths and weaknesses of the GBMA is (to a large extent) presented in Section 5.2. Furthermore, we remark that small-data modeling is a new development in comparison with (1) the generic WPM framework [Englehardt et al., 2014] and (2) our extended web-tracking model [Van Eijk, 2012]. By modeling the context with specific small-data elements we will later address RQ2 (see M-step 1 and M-step 2, Subsection 4.4.3). At this moment we will focus on RQ1.

3.5 AN ANSWER TO RESEARCH QUESTION 1

We are now able to answer RQ1: how do we move from data collection to graph analysis? To answer this question we presented a Graph-Based Methodological Approach. Through the GBMA we showed that small-data models can be based on metadata.

Since the HTTP header contains valuable information needed to build small-data models for web tracking, we thoroughly examined data collection (Section 3.1) and data reduction (Section 3.2). Building models is only possible with sufficient knowledge of data modeling (Section 3.3).

Our findings were extensively explained and thoroughly tested. We arrived at the following path that straightforwardly goes from data collection to graph analysis in three well defined parts.

FINDING 1: The properties of the nodes and edges in the small-data model correspond with metadata in the research data. For instance, the edges in the small-data model may be directed or weighted, to express specific properties of the metadata.

FINDING 2: The main improvement of the capabilities for analysis is the fact that a small-data model allows us to apply a variety of algorithms to the data.

FINDING 3: Such a varied algorithmic approach allows us to deepen our understanding of what actually happens when we visit websites containing RTB advertisements. This is the answer to RQ1.

The new and deep insight into the intricacies of web tracking is important for science, but also for society, since it substitutes a warning to privacy protections: be aware of the power of modern technologies.

Categorization of RTB systems

This chapter addresses RQ2: what are the emerging characteristics of the graph that is fit for graph analysis? The research question serves to improve our understanding of RTB systems.

The theoretical (marketing) models provided by companies often present a limited conceptualization of what actually happens. In fact, companies do not really have an interest in explaining the details of how their technology works. Moreover, if the description of the technology becomes more detailed, it does not necessarily mean it is better for the understanding of the end-user. Instead, they are left with a description of a black box. Marco Kloots is the CEO at Platform161. With a quote, he explained the black box in an interview by AdEchanger:

„The more third-party tech an advertiser is connected to, especially tech that is data-related, the greater the chance that they are not well enough equipped to explain what exactly happens all the way down the line. It would be difficult to offer a clear view of the data they are using and exactly where it comes from.”²⁴⁶

Whatever the case, we can at least peek into the black box because the traces of the data of the websites we visit contain all sorts of metadata. So, the real question is: what is happening in the black RTB box? It turns out that when we apply different algorithms that have been widely used in network science to the RTB context, we learn about the data (and therefore which companies) are connected to RTB systems.

²⁴⁶ URL: <https://adexchanger.com/data-driven-thinking/gdpr-horizon-data-challenges-opportunities-loom/> (29 September 2017).

To reach our goal of answering RQ2, we construct a specialized graph model from the metadata. The graph model reflects the interactions between the end-user visiting websites with ads and companies collecting the data enabling the ads. The graph model lets us categorize RTB systems and provide an answer to the question who are the most influential actors? If we look at the metadata in that way, then characteristics of the graph emerge which enables us to distinguish RTB systems.

In Section 4.1 we investigate RTB systems as a network of partners. In Section 4.2 we arrive at a definition for RTB system. In Section 4.3 we give a *theoretical* view of the key concepts in RTB systems reflected in partner networks. After the theoretical view, we present an *empirical* view (Section 4.4) and a *legal* view (Section 4.5) of RTB deployed on national and regional European news websites. Then, we discuss our graph analysis by applying network science algorithms to our contextual research data (Section 4.6). It is the stepping stone to understand precisely how we categorize RTB systems. In this section we apply the GBMA to graph analysis of partner networks in RTB systems within the context of news websites in Europe. We are encouraged to follow this approach since the application of graph measures to data has been successful in other research fields. For instance, (1) Barabási [2016] applied network science algorithms to spreading processes on networks, e.g., to quantify and forecast the spread of infectious diseases,²⁴⁷ (2) Guye, Bettus, Bartolomei, and Cozzone [2010, p. 188] applied network science algorithms to neuroscience, (3) Wagner and Neshat [2010, pp. 124–125] assessed the vulnerability of supply chains using graph theory, and (4) Dunn, Dudbridge, and Sanderson [2005, p. 11] are known for their application of graph theory to bioinformatics. We end this chapter with conclusions (Section 4.7) and we provide an answer to RQ2 (Section 4.8).

Since the technological development is overwhelming and the results are clearly disruptive, we give the reader now and then a break by headers, such as ‘progress of developments’ and ‘progress of discussion’. Moreover, we complete each section with sec-

²⁴⁷ Barabási [2016, 34–41] introduced „a network based approach to epidemic phenomena that allows us to understand and predict the true impact of these hubs. The resulting framework, that we call network epidemics, offers an analytical and numerical platform to quantify and forecast the spread of infectious diseases.” Hubs are defined as nodes with an exceptional number of links in the contact network on which a disease spreads.

tion conclusions. In this way, we hope to guide the reader adequately in this fascinating world.

Our main finding is that the roles of the companies involved with web tracking can be explained through expressing metadata from the HTTP header into (mathematical) characteristics of the nodes and edges of the graph. The finding corresponds with Step 6 of our research methodology, viz. analysis of the research data with intelligent techniques.

4.1 THE NETWORK OF PARTNERS

Van Eijk and Chester [2014; 2015] investigated RTB systems by zooming in on acquisitions, mergers, and strategic partnerships of companies specializing in RTB technology over at least a period of three years. We composed a table to reflect the consolidation (henceforth: consolidation table). The consolidation table is published over seven pages (i. e., Table C.1 to Table C.7).

The consolidation table shows 443 acquisitions and mergers by 58 companies. The companies are alphabetically ordered. In square brackets is mentioned which acquisition or merger took place.²⁴⁸ Building a network of partners is the key to develop an own category of RTB systems. The intriguing question is: how many categories will there be in 2020, 2030 and so on? In fact, we would like to know the trend. Therefore, we describe the situation in December 2015, and more importantly the trend over the period 2013 - 2015 (three years). In 2015, Google was leading with 65 acquisitions/mergers followed by Yahoo (50), Twitter (27), Facebook (25), and Oracle (24).²⁴⁹

As stated above, the observation of the consolidation took place between January 2013 and December 2015. It would take too long to describe them all one by one. Instead, in Subsection 4.1.1 we provide brief descriptions of two acquisitions, (A) the comScore/Proximic *acquisition* and (B) the Dunnhumby/Sociomantic *acquisition*. In Subsection 4.1.2 we see that the latter acquisition led to (C) the *merger* between Kroger and Dunnhumby and (D) the Rentrak-comScore *merger*.²⁵⁰ In Subsection 4.1.3 four *strategic partnerships*

²⁴⁸ The mergers were verified by manual Google searches. The acquisitions in our table were verified by manual lookups in the <http://crunchbase.com> database (7 August 2016).

²⁴⁹ See also Section 2.1.

²⁵⁰ Please note that Kantar/WPP invested \$ 244 million in comScore as part of a strategic partnership instead of an acquisition. URL: <https://www.comscore.com/>

are outlined: (E) Microsoft, (F) Facebook, (G) Google, and (H) PayPal. In Subsection 4.1.4 we provide an example of an RTB category: (I) The Rubicon Project, Inc. (henceforth: the Rubicon Project). Finally, in Subsection 4.1.5 we provide section conclusions.

4.1.1 *Acquisitions*

Below we describe two acquisitions, in which the possession of real-time bidding technology played a key role.

A: COMSCORE/PROXIMIC [ACQUISITION]

comScore acquired ProximiC because it owned real-time bidding technology allowing them to reach an audience by creating customized targeting segments using metadata of the website that an end-user is visiting. They did so by using, e. g., keywords, phrases, or sentences to give meaning to the context in real time.²⁵¹

B: DUNNHUMBY/SOCIOMANTIC [ACQUISITION]

Dunnhumby showed a special interest in Sociomantic for its (1) real-time ad technology and (2) tracking data from more than 700 million online consumers. With the acquisition Dunnhumby hoped to fulfill its goals to improve „how advertising is planned, personalized and evaluated”.²⁵²

Insights/Press-Releases/2015/2/comScore-and-Kantar-Announce-Strategic-Global-Partnership (10 October 2015). Kantar/WPP's investment has not been added to the table.

²⁵¹ The press release reads as follows: „Using ProximiC's solutions, buyers can create customized targeting segments using contextual data, brand protection features, keywords, phrases or sentences to target and filter the inventory that is being bid on in real time. (...) ProximiC's solutions are currently integrated into several publishers and demand-side platforms (DSPs), including AppNexus.” URL: <http://www.comscore.com/Insights/Press-Releases/2015/5/comScore-Acquires-Proximic-to-Bolster-Pre-Bid-Solutions-for-Buyers-and-Sellers> (8 December 2015).

²⁵² The press release states: „Dunnhumby will combine its extensive insights on the shopping preferences of 400 million consumers with Sociomantic's intelligent digital-advertising technology and real-time data from more than 700 million online consumers to dramatically improve how advertising is planned, personalized and evaluated.” URL: <https://www.dunnhumby.com/dunnhumby-acquires-sociomantic-revolutionise-digital-advertising> (9 December 2015).

4.1.2 Mergers

Below we provide two brief descriptions of mergers, that were in fact direct consequences of the acquisitions described above.

C: KROGER/DUNNHUMBY [MERGER]

After the acquisition of Sociomantic, the strengthened position of the company became of crucial importance for Kroger.²⁵³ Dunnhumby merged with the relatively larger Kroger, forming a new wholly-owned subsidiary.²⁵⁴ A key role for Dunnhumby was to retain staff critical to strategic innovations in customer science.²⁵⁵

D: RENTRAK/COMSCORE [MERGER]

Rentrak merged with comScore because they owned specialized technology allowing them to track the end-user online whenever and wherever they visited a website and played a video. Combining the technological capabilities of both companies comScore is able to provide a more complete picture of the way people engage with online video, online movie, and digital television.²⁵⁶

4.1.3 Strategic partnerships

In addition to the phenomenon of acquisitions and mergers, companies are looking for *strategic partnerships* to create a network

²⁵³ The press release states the following: „Dunnhumby will now have the ability to use its proven insight products and data expertise to capture the substantial, previously unavailable potential of the North American market through working with new retailers, consumer brands and media partners.” URL: <https://www.dunnhumby.com/dunnhumby-ltd-and-kroger-announce-new-relationship> (9 December 2015).

²⁵⁴ Ibid, „More than 500 of dunnhumbyUSA’s employees will become associates of 84.51°, a wholly-owned subsidiary of The Kroger Co.”

²⁵⁵ Ibid, „Dunnhumby will continue to pioneer the field of customer science through innovations in retail consulting services, analytics software, data science and digital media and will continue to develop its existing strong platform of client relationships in the US.”

²⁵⁶ The strategic rationale provided in the press release is as follows: „Together, comScore’s industry-leading digital audience and advertising solutions, combined with Rentrak’s census-based worldwide movie and video-on-demand measurement, and its massive and passive TV measurement offerings, will provide a more complete picture of the way people consume media today and in the future.” URL: <http://www.comscore.com/Insights/Press-Releases/2015/9/comScore-and-Rentrak-to-Merge> (8 December 2015).

of connections. Below we provide four prominent examples: (E) Microsoft, (F) Facebook, (G) Google, and (H) PayPal.

E: MICROSOFT

Microsoft invested strategically in the technology stack for real-time bidding owned by AppNexus. They did so since they lacked the technology that AppNexus had. So far, Microsoft's behavior was unlike its competitors - e. g., Yahoo, Amazon, AOL, Facebook, the Rubicon Project and Google - since Microsoft did not have an own real-time bidding technology (cf. Shields [2014]).

F: FACEBOOK

Facebook introduced real-time bidding in 2012, i. e., Facebook Exchange. The RTB technology relies on cookies. Cookies work primarily for desktop (re)targeting. They can be categorized into two types: (1) a cookie containing a unique identifier which indicates whether an end-user is logged in to Facebook and (2) a cookie containing the browser's history. The latter cookie contains valuable information for companies that partner with Facebook. It contains the interaction of an end-user with the company's marketing campaign on its website on facebook.com. In addition, Facebook brings in *if* and *when* an end-user is on Facebook (cf. Smith [2014]).

G: GOOGLE

Google shows an extreme expansion drift. For instance, its certified vendors network for the DoubleClick RTB exchange contained 836 companies on 20 January 2015, and 2,106 companies on 8 December 2015.²⁵⁷

²⁵⁷ DoubleClick serves third-party ads via its ad-exchange platform. „The certified vendors permitted to make 3rd-party calls are generally of the following type: Demand Side Platform, Agency Trading Desk, Ad Network, Ad Exchange, Standard Ad Server & Rich Media Vendors. The certified vendors permitted to make 4th-party calls are generally of the following type: Research products, which include Analytics/Performance, Brand-Lift Studies, & Verification Services.” URL: <https://support.google.com/3pasertification/table/4570113?hl=en> (8 December 2015).

H: PAYPAL

Recently, PayPal disclosed a list of third parties (other than PayPal customers) with whom the company shares data [PayPal, 2017; 2018].²⁵⁸ Google is one of its partners. The added value for PayPal is that Google's technology enables them:

„to help identify behaviour on PayPal websites and the mobile app in order to guide decisions about *targeted marketing*; to help efficiently handling and optimising desktop and mobile campaigns and elsewhere in the web and [to] execute retargeting campaigns in order to deliver *personalised advertising*.” [PayPal, 2018] (emphasis added)

To fulfill the purposes of targeted marketing and personalized advertising, PayPal shares the following nine types of metadata with Google (cf. PayPal [2018]):

- (1) UID generated by (a) cookies, (b) pixel tags, or (c) similar technologies embedded in webpages, ads and emails delivered to end-users,
- (2) advertising ID,
- (3) device ID,
- (4) encrypted e-mail address,
- (5) customer ID,
- (6) merchant ID,
- (7) transaction value,
- (8) transaction ID, and
- (9) loan approval amount.

4.1.4 The Rubicon Project as an example

We performed a longitudinal analysis on the Rubicon Project's RTB file 'emily.html'.²⁵⁹ The RTB file contains metadata about the

²⁵⁸ PaypPay disclosed it first list on 1 October 2017 [PayPal, 2017] and published an updated version on 1 January 2018 [PayPal, 2018].

²⁵⁹ The Internet Wayback Machine has been archiving the Rubicon Project's emily.html file since 2011. The file had been retained 364 times between 23 March 2011 and 7 December 2016. URL: https://web.archive.org/web/*/http://tap2-cdn.rubiconproject.com/partner/scripts/rubicon/emily.html (26 March 2018). I retained a copy of the files and processed them with common command line tools, i. e., „grep 'rtb_sync =' 120224 emily.html >120224 emily.json” and counting the number of instances, i. e., „grep 'partner:.' 120224\emily.json | wc -l”.

Rubicon Project’s partner network. Our analysis includes the following four elements: (1) partner, (2) image, (3) iFrame, and (4) script. In Figure 4.1 we give the results of our analysis of the partner network of the Rubicon Project between March 2011 and December 2016 (almost six years).

The figure shows the total number of references per metadata element in the file ‘emily.html’ to the Rubicon Project’s partners (denoted as Partner, depicted by a blue line), image tags (denoted as Image, depicted by a red line), iFrames (depicted by an orange line), and JavaScripts (denotes as Script, depicted by a green line). The number of references per metadata element in December 2016 is as follows: partner (97), image (149), iFrame (6), and script (3).

The analysis shows that only a few partners are referenced with a script tag or an iFrame. The image tag is the preferred mechanism to reference a partner within the Rubicon Project RTB system.²⁶⁰

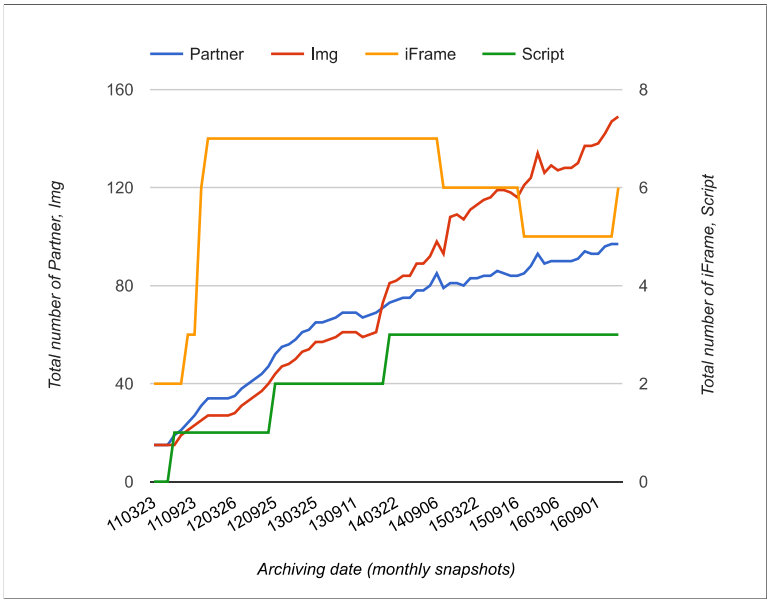


Figure 4.1: Longitudinal analysis of the Rubicon Project partner network between March 2011 and December 2016.

²⁶⁰ The image tag is a web beacon. See, e. g., AppNexus [2017b].

PROGRESS OF DEVELOPMENTS

I presented and discussed the results to the Article 29 Data Protection Working Party (Art. 29 WP) [Van Eijk, 2016a] and at the Dagstuhl seminar 17162, Online Privacy and Web Transparency [Van Eijk, 2017]. The outcome of the discussions was that the increase in RTB partners in the Rubicon Project partner network is similar to the increase - during the same period - in the number of HTTP cookies in the browser.²⁶¹ The audience agreed with the trends and the conclusions (see Subsection 4.1.5).

4.1.5 Section conclusions

Three new marketing technologies, viz. *real-time* analytics, *real-time* data attribution, and *real-time* (algorithmic) bidding, intensified no less than four issues, viz. (1) the tracking of user preferences, (2) demographics, (3) geolocation, and (4) user behavior on the web. In passing we mention that the list of new marketing technologies is much longer. Speaking in general, we may usually distinguish two developments: (a) advancement of tracking technologies and (b) an increase of the number of partners participating in an RTB system.

To summarize, we may conclude that (1) each network of partners relies on its own targeted-advertising framework and (2) has its unique way of tracking end-users and bidding for end-users.²⁶² Typical RTB tasks that partners perform are:

- (1) referrals,
- (2) web analytics,
- (3) audience segmentation,
- (4) personalization, and
- (5) (re)targeted advertising.

What partners in an RTB system do is strengthening their positions. The combined new network allows companies to

- (1) expand their client portfolio,
- (2) reach (new) online prospects and consumers,

²⁶¹ Figure 4.1 confirms observations in the initial period of RTB (between the years 2010 and 2013) of Van Eijk [2011b] on the increasing number of RTB partners (see Kamphuis [2013b] and Kamphuis [2013a]).

²⁶² Cf. Fielding [2015].

- (3) gain tracking data about end-users,
- (4) prevent tracking data leaking to competitors, e. g., from their cookie pool, and to some extent
- (5) gain access to (third-party) ad technology.

All in all, our final section conclusion is that the developments are fast, disruptive and have big impact on the society.

4.2 A DEFINITION FOR RTB SYSTEM

At the end of the above observations, we are still facing the task to formulate a definition for an RTB system. If we combine RTB as a technology (Definition 1.1) with the network of partners, the description of an RTB system is as follows.

DEFINITION 4.1: A REAL-TIME BIDDING SYSTEM is defined to be a network of partners enabling big data applications within the organizational field of marketing to improve sales by *real-time* data-driven marketing and personalized (behavioral) advertising.

We remark, that our definition is primarily aimed at web tracking for *commercial* purposes. Earlier on (see Subsection 2.1.1), we mentioned the secondary use of RTB for other purposes, i. e., State tracking. In our discussion this holds for the question why WPM matters?

4.3 THEORETICAL VIEW OF RTB SYSTEMS

In Figure 4.2 we give a schematic picture of the theoretical view of the current state of technology of RTB systems.²⁶³ A clear understanding about specialized terms and key components of RTB systems is a prerequisite for the classification of RTB systems.²⁶⁴ Other theoretical views have been provided by, e. g., Zhang, Yuan, Wang, and Shen [2014], Zhang, Yuan, and Wang [2014], Olejnik and Castelluccia [2016], Ryan [2017, p. 21], and Papadopoulos, Rodriguez, Kourtellis, and Laoutaris [2017, p. 2].

²⁶³ For a historical overview we refer to MIT Technology Review Custom [2013] and IABureau [2017b, p. 3].

²⁶⁴ I drew inspiration from a model provided by VertaMedia. URL: https://vertamedia.com/assets/upload/content/admarket_integration.png (24 September 2017).

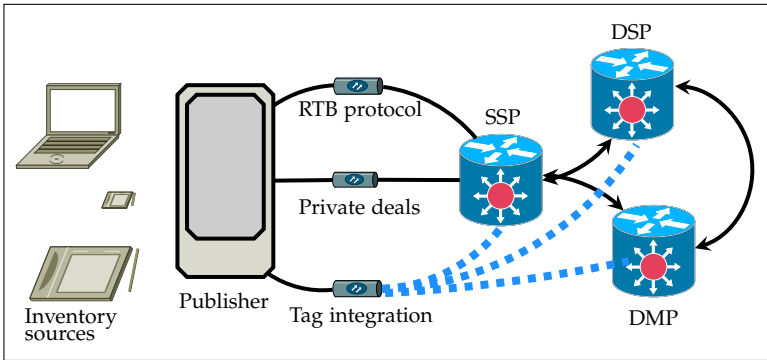


Figure 4.2: Theoretical view of RTB systems.

Of course, we understand the risk of oversimplification, but we believe it is beneficial (a kind of courtesy) to the reader to provide an adequate and relevant background. For this background we introduce eight key building blocks used in RTB systems. We define them at the beginning of the subsection and briefly discuss each of them:

- (1) inventory sources (Subsection 4.3.1),
- (2) publisher (Subsection 4.3.2),
- (3) Supply Side Platform (SSP) (Subsection 4.3.3),
- (4) Demand Side Platform (DSP) (Subsection 4.3.4),
- (5) Data Management Platform (DMP) (Subsection 4.3.5),
- (6) RTB protocol (Subsection 4.3.6),
- (7) private deals (Subsection 4.3.7), and
- (8) tag integration (Subsection 4.3.8).

Once more, we reiterate that the composition of the eight building blocks is depicted in Figure 4.2. The interaction between the blocks is described in the eight subsections below. We close this section with a brief overview of the nine steps in the RTB-bidding process (Subsection 4.3.9) and with section conclusions (Subsection 4.3.10).

We anchor our theoretical view in specialized terms and mark that RTB is known for its frequent use of subject terms.²⁶⁵

²⁶⁵ For a description of *general* ad-tech terms we refer to, e.g., the glossary of AppNexus [2017b]. For *specialized* terms we refer to four glossaries, e.g., (1) IABureau Audio Council [2017], (2) IABureau Mobile Advertising Council [2017], (3) IABureau Video Advertising Council [2017], or (4) IABureau Ad Effectiveness Council [2017].

4.3.1 *Inventory sources*

DEFINITION 4.2: An INVENTORY SOURCE is the screen of an end-user’s (mobile) device. The screen can display one or more ad spaces that a publisher has available to sell to an advertiser.

Figure 4.2 illustrates the laptop, phone, and tablet as inventory sources. The devices vary in screen size, which is an important feature for display advertising. The screen size is an indication of the *viewable* size of the ad. Other examples of inventory sources are (a) digital signage in public places and (b) personalized price tags in your local supermarket (in the near future).

Table 4.1: Analysis of distinct ad-slot sizes.

WIDTH [PIXELS]	HEIGHT [PIXELS]	AD-SLOT TYPE	COUNT	PERCENT
300	250	medium rectangle	11,599	54.7 %
728	90	leaderboard	6,299	29.7 %
160	600	skyscraper	2,116	10.0 %
320	50	mobile-optimized banner	1,203	5.7 %
300	90	mobile-optimized banner	1	0%
301	250	medium rectangle	1	0%
			21,219	

Table 4.1 shows the results of our investigation into ad sizes. We analyzed a configuration file (adInfo.js) of the Brave browser [Eich, Bondy, et al., 2016].²⁶⁶ The file contains metadata about ad slots, i.e., the height, the width, and network identifier of 21,219 ad slots.²⁶⁷

Moreover, Table 4.1 contains an overview of the distinct ad size (in pixels) of the ad space available to advertisers. From the Brave data, we find that the so called medium rectangle advertisement

²⁶⁶ The Brave browser [Eich et al., 2016] is based on a revolutionary business model that aims to share the money made with ads with end-users. To keep that promise, the browser takes control of the ads replaces them with ads sold by Brave Software Inc.

²⁶⁷ Source: URL: <https://github.com/brave/browser-laptop> (4 July 2016, tag: vo.10.3dev).

is leading with 54.7% followed by the leaderboard (29.7%) and the skyscraper (10.0%).

PROGRESS OF DISCUSSION

I presented these results to the Art. 29 WP Technology Subgroup [Van Eijk, 2016a]. The outcome led to a better understanding of how RTB works, i.e., by differentiating between the ad slot and other RTB technologies, such as measuring the ad.

4.3.2 *Publisher*

DEFINITION 4.3: „A PUBLISHER is a person or corporation whose business is publishing.” [Merriam-Webster.com, n.d.]

Many digital publishers, e.g., news websites, depend on RTB for their business model. They do so by selling ad space on their websites to media buyers and advertisers. The aim is to serve (personalized) ads to the screen of unique (recurring) visitors. Obviously, media buyers and advertisers are willing to pay more if the ad it is viewed by the intended audience. To reach that goal, information is collected about the audience, e.g., usage and performance data.

4.3.3 *Supply Side Platform*

DEFINITION 4.4: A SUPPLY SIDE PLATFORM (SSP) enables publishers to *auction* their ad slots to all media buyers and advertisers. An SSP specializes in matching advertisers with the SSP’s publisher network.

Bidding strategies for a SSPs have been studied by, e.g., Balseiro, Feldman, Mirrokni, and Muthukrishnan [2014] and B. Chen, Yuan, and Wang [2014]. An example of a well-known SSP is Improve Digital, an SSP based in the Netherlands.²⁶⁸ Improve Digital specializes in more than 40 RTB system integrations. In this way they connect about 80,000 advertisers to over 3,500 media buying part-

²⁶⁸ Improve Digital is a partner of Turn. URL: <http://www.improvedigital.com> (6 August 2016).

ners. Other examples of SSPs are AppNexus,²⁶⁹ Pubmatic, and the Rubicon Project.

PROGRESS OF DEVELOPMENT: HEADER BIDDING

DEFINITION 4.5: **HEADER BIDDING** enables publishers to add rules for an ad auction.

To be complete, we mention a recent technical improvement: header bidding [IABureau, 2017d]. The technology enables publisher to reach out to an ad exchange immediately after an end-user requests a webpage, to ask for bids (cf. IABureau [2017a]). Header bidding usually applies to a specific ad on a publisher's website, i.e., the leaderboard (see Table 4.1). Header bidding enhances the strategic bidding capabilities of publishers (cf. Jauvion, Grislain, Dkengne, Garivier, and Gerchinovitz [2018]).²⁷⁰

The partnership between OpenX and Google (see, e.g., Fairchild [2018] or Olejnik and Castelluccia [2016]) is a very recent example of an implementation of header bidding. OpenX partnered with Google to offer its publishers header bidding. The company enabled its SSP with Google Exchange Bidding, the header bidding service that Google offers to its DoubleClick partners.²⁷¹

4.3.4 *Demand Side Platform*

DEFINITION 4.6: A **DEMAND SIDE PLATFORM** enables its network partners to *bid* for ad slots. Media buyers and advertisers bid based on criteria such as, (geo)location, gender, browsing history.

A DSP specializes in running an advertising campaign on different websites while targeted at the intended audience at the right time. Bidding strategies for DSPs have been studied by, e.g., Vines, Roesner, and Kohno [2017], Zhang, Yuan, and Wang [2014], Zhang, Yuan, Wang, and Shen [2014], and Wang et al. [2016]. An example of a DSP is iPinYou which connects about 1,800 companies with

269 AppNexus Publisher SSP. URL: <https://www.appnexus.com/en/publishers/publisher-ssp> (9 August 2016).

270 See also, e.g., Y. Chen [2017], Wang, Yuan, and Cai [2015], Wang, Yuan, and Zhang [2016], or Grigas, Lobos, Wen, and Lee [2017].

271 A second recent example is the (technical) implementation of header bidding technology by Media Impact [2018, p. 1].

well-known brands to its RTB systems. Other examples of DSPs are Adchemy,²⁷² AppNexus,²⁷³ Adform, DataXu, Facebook, Gravity4, MediaMath, TubeMogul, Turn, Videology, and [X+1].²⁷⁴

PROGRESS OF DEVELOPMENT: BIDDING-AS-A-SERVICE

DEFINITION 4.7: BIDDING-AS-A-SERVICE is a cloud-based technology connecting advertisers to multiple RTB systems and enabling them to respond to RTB-bid requests in real-time.

A recent development is bidding-as-a-service. Below, we provide two examples: Beeswax [Beeswax, 2017] and Open Bidder [Google, 2016b]. Beeswax runs as a service on Amazon Web Services (AWS). Open Bidder runs as a service on Google Cloud Platform. Both frameworks give advertisers control over ad pricing in real time.²⁷⁵ The services enable advertisers to respond in real-time to bid requests, viz. step 4 in the RTB-bidding process (infra Subsection 4.3.9). Moreover, the services enable them to place bids on an increasing number of exchanges due to the support for a variety of common RTB protocols (see Table 4.2).²⁷⁶ Therefore, an advertiser bidding via these services does not have to implement each common RTB protocol herself. Instead, implementing the bidding-as-a-service's API suffices.

4.3.5 Data Management Platform

DEFINITION 4.8: A DATA MANAGEMENT PLATFORM (DMP) enables DSPs and SSPs to *zoom in* on their audience. A DMP specializes in customer data.

For a DSP having access to customer data means that it puts them in a better position to (re)target the intended end-user (audience)

²⁷² Acquired by Walmart Labs (Table C.1).

²⁷³ AppNexus Programmable DSP. URL: <https://www.appnexus.com/en/agencies-and-advertisers/programmable-dsp> (8 August 2016).

²⁷⁴ Acquired by RocketFuel (Table C.1).

²⁷⁵ URL: <https://docs.beeswax.com/docs/beeswax-architecture-life-of-a-bid> (29 January 2018). The Beeswax bidding API enables advertisers to handle ad-campaign decisions in real-time, e.g., (1) „Does targeting match the incoming request? (2) Is the user over the desired frequency cap? (3) Does the ad group have budget available? and (4) Does the creative associated with the ad group match any creative attributes in the incoming request?“ (internal examples omitted)

²⁷⁶ See, e.g., URL: <https://docs.beeswax.com/docs/release-notes> (29 January 2018).

in a specific context. For an SSP having access to customer data means that it puts them in a better position to personalize the content on *their* website offered to end-users.

An example of a well-known DMP is Bluekai:²⁷⁷ „With more than 30 branded data providers for 3rd party data, marketers have access to nearly 700 million anonymous customer profiles and 40,000 data attributes.”

PROGRESS OF DEVELOPMENT

Other examples of DMPs are Adobe Audience Manager, Krux, and Relay42.²⁷⁸ Moreover, we remark that the previously mentioned DSPs Turn, and [X+1] also specialize in DMP technology.

4.3.6 RTB protocol

DEFINITION 4.9: An RTB PROTOCOL is a standard used to define a method of exchanging bidding data, e. g., bid requests, over an RTB system.

Table 4.2: Overview of eight common RTB protocols.

OWNER / AUTHOR	RTB PROTOCOL	PUBLIC
AppNexus	Creative API	✓
Facebook	Facebook Exchange	(not publicly available)
Google	DoubleClick Adx API	(not publicly available)
Google	OpenRTB (protol buffer)	✓
Google	OpenRTB (JSON)	✓
IPONWEB	BidSwitch Protocol	✓
IABureau	OpenRTB	✓
IABureau	OpenDirect	✓

Table 4.2 shows two proprietary and six publicly available RTB protocols provided by five organizations. It would take too long

²⁷⁷ Acquired by Oracle (Table C.1). acsURL: <https://www.oracle.com/marketingcloud/products/data-management-platform/index.html> (9 August 2016).

²⁷⁸ For an example of the capabilities of Krux, see Subsection 2.3.2.

to describe them all one by one. Instead, we provide an example of an OpenRTB-bid request [IABureau, 2017b] below.²⁷⁹

AN EXAMPLE OF AN OPENRTB BID REQUEST

In Listing 4.1 we provide an example of an OpenRTB [IABureau, 2017b] bid request.²⁸⁰

Listing 4.1: Example RTB bid request.

```

1 POST /auctions HTTP/1.1
2 Content-Type: application/json
3 Content-Length: 640
4 accept: */*
5 connection: Keep-Alive
6 x-openrtb-version: 2.3
7 (...)
8 "user": {
9     "id": "23456",
10    "buyeruid": "202122",
11    "data": [{
12        "id": "303132",
13        "name": "Data Provider X",
14        "segment": [{"name": "online news"
15    }, {
16        "id": "505148",
17        "name": "data-X-location",
18        "value": "Midwest USA"
19    }, {
20        "id": "505152",
21        "name": "data-X-age",
22        "value": "40-50"
23    }, {
24        "id": "404142",
25        "name": "data-X-buying-intent",
26        "value": "high" }]
27    }]
28 }
```

²⁷⁹ Cf. CM Summit and BattelleMedia [2013]: Behind the banner, a visualization of the ad-tech ecosystem.

²⁸⁰ OpenRTB protocol (version 2.3).

The example demonstrates the use of metadata in a bid request. We modeled the bid request after the end-user in the short video ‘Behind the banner’ [CM Summit & BattelleMedia, 2013]: „I am a 32-year-old working mother, living in the Midwest of the USA. I usually browse Etsy and Wikipedia but now I am reading a New York Times article. Oh, and I just looked at a new pair of glasses on Ebay.”²⁸¹

PROGRESS OF POSSIBLE OBSTACLES

In closing, concerning RTB protocols, we remark that interconnection becomes a major bottleneck when network partners aim for integration with as many RTB systems as possible (cf. BidSwitch [2016]). It may result in hindering RTB from more advanced technological progress.²⁸² We remark that standardizing protocols - i. e., bidding-as-a-service APIs (see subsection 4.3.4) - may be a win-win strategy for all network partners involved.

4.3.7 *Private deal*

DEFINITION 4.10: A PRIVATE DEAL is a specialized contract between a publisher and selected advertisers.

The technology enabling private deals is SSP technology (Supply Side Platform). Private deals are beneficial for publishers. For instance, a brand may have an interest in buying inventory for a special price on websites it believes its buyers are present.²⁸³ Other terms used for private deals are, e. g., publisher-direct deal, preferred deal, private auction, or deal ID [IABureau, 2015].

²⁸¹ The bid request is formatted as an HTTP POST request (Listing 4.1, spanning rr. 1-6) with a JavaScript Object Notation (JSON) payload. Furthermore, the end-user was tagged with the (fictitious) ID value ‘23456’ (Listing 4.1, spanning r. 9) as an active middle aged runner (Listing 4.1, spanning rr. 14-25). The advertiser’s buyer ID is listed in row 10. The data array shows the ID and the name of the DMP (Listing 4.1, spanning rr. 12-13).

²⁸² BidSwitch [2016]: „The direct integration process, both commercially and technically, becomes a major bottleneck to scaling trading activity and growing revenues. The race to integrate with as many partners as possible is dramatically hindering the ecosystem from more advanced technological progress.”

²⁸³ AppNexus [2017c]: „Typically, to initiate the purchase and sale of deals and packages, a *publisher* invites an advertiser to bid on its inventory, and it enables the advertiser to gain first access on specific ad inventory before it’s made available to other buyers in an open auction.” (emphasis added)

PROGRESS OF DEVELOPMENT: DEAL ID

Deal ID is a recent technical improvement. These technologies enable publishers to monetize ad-slots directly with media buyers and advertisers through private deals.²⁸⁴

4.3.8 Tag-Based Integration

DEFINITION 4.11: „TAG-BASED INTEGRATION (TBI) is the use of specialized HTML code to simplify the collection of event-tracking data, e. g., clicks by the end-user, tracking impressions, and conversions." [AppNexus, 2017b] (slightly modified)

We remark that Tag-Based Integration includes the three technologies visualized in Figure 4.1, i. e., (a) iFrame, (b) script, and (c) image tag (beacon technology) (see, e. g., C. Thomas, Kline, and Barford [2016]).²⁸⁵ Furthermore, we briefly addressed the use of tags and the KLM case study against the background of event tracking (see Subsection 2.3.1 and Subsection 2.3.2).²⁸⁶ KLM integrates:

- (1) the Relay42 tag [Relay42, n.d.],
- (2) the Google Floodlight [Google, n.d.-c] tag, and
- (3) Google analytics event data (see Subsection 2.1.3).

The combination of the two tags - a real-time *data-attribution* tag and a real-time *analytics* tag - with Google analytics event data allows KLM to track end-users across their marketing channels online and communicate with them through RTB.²⁸⁷ The specializations in *tag management* and in *event tracking* explain why the Relay42 tag and the Google Floodlight tag put KLM in a good position to transform event-tracking data to end-user knowledge (viz. Section 1.3).

²⁸⁴ AppNexus [2017a]: „It is best practice to operate on a flat bidding structure where the base CPM is at least 2x higher than the pre-negotiated floor price. This ensures delivery and scale on the deal ID.”

²⁸⁵ In contrast to the '1x1'-pixel, a 'oxo'-svg image may be used by websites. See, e. g., URL: https://www.avrotros.nl/typo3conf/ext/www_resources/Resources/Public/GFX/watermerk.svg (20 September 2018).

²⁸⁶ See Subsection 2.3.2.

²⁸⁷ See n. 92. [Google, n.d.-c]: „A Floodlight activity is a specific conversion you want to track, such as the completion of a purchase or a visit to a page on your site. A tag is automatically generated for each activity, and your web team installs the tag onto your site. When a visitor lands on the conversion page, the tag reports a conversion.”

PROGRESS OF DISCUSSION

I presented and discussed the KLM case study and our theoretical view of RTB system at the annual privacy conference Computers, Privacy, and Data Protection [Van Eijk, 2018, p. 4, p. 10]. The outcome of the discussion was that the deployment of Tag-Based Integration is not limited to DMPs only. Nowadays, SSPs and DSPs rely also on tag-based technology.²⁸⁸ This insight is reflected with a blue dotted line in our theoretical view (see Figure 4.2).

4.3.9 *Nine steps in the RTB-bidding process*

DEFINITION 4.12: „The RTB-BIDDING PROCESS is a series of actions or operations conducing to the delivery and verification of an online advertisement.” [Merriam-Webster.com, n.d.] (slightly modified definition of a process)

Below we provide a detailed list of the nine steps (called RTB-steps) in the RTB-bidding process (cf. Ryan [2017, pp. 18, 21]).

RTB-STEP 1: An end-user requests a webpage.

RTB-STEP 2: The publisher’s ad server on the webpage selects an SSP.

RTB-STEP 3: The SSP selects an ad exchange.

RTB-STEP 4: The ad exchange sends bid requests to *hundreds* of network partners and enables them to generate a bid response.²⁸⁹

RTB-STEP 5: The ad exchange permits preferred DMPs/DSPs to synchronize HTTP cookies.²⁹⁰

RTB-STEP 6: The ad exchange serves the winning bid.

RTB-STEP 7: The DSP serves the ad-agency’s ad.

RTB-STEP 8: The ad loads from a CDN.

RTB-STEP 9: The advertising agency’s ad sever loads a verification JavaScript.

These RTB-steps illustrate the interplay between SSP, DSP, and DMP. The interplay is visualized in a sequence diagram in Fig-

²⁸⁸ I am indebted to Chester [2018] and Polonetsky [2018] who confirmed this insight independently from one another.

²⁸⁹ See also Figure 4.1 and text to n. 261.

²⁹⁰ See, e.g., URL: <https://docs.beeswax.com/docs/cookie-syncing> (39 January 2018).

ure 4.3.²⁹¹ We remark that some companies specialize in multiple building blocks. For instance, the Rubicon Project specializes in SSP and DSP technology (see Subsection 4.1.4).²⁹² Furthermore, we remark that the request in RTB-step 1 is not constrained to a request from a webpage. The request could be made from a variety of inventory sources (see Subsection 4.2) and big-data applications (see Figure 1.1).

Before we turn to the section conclusions, I would like to give thought to the term *ad exchange*.

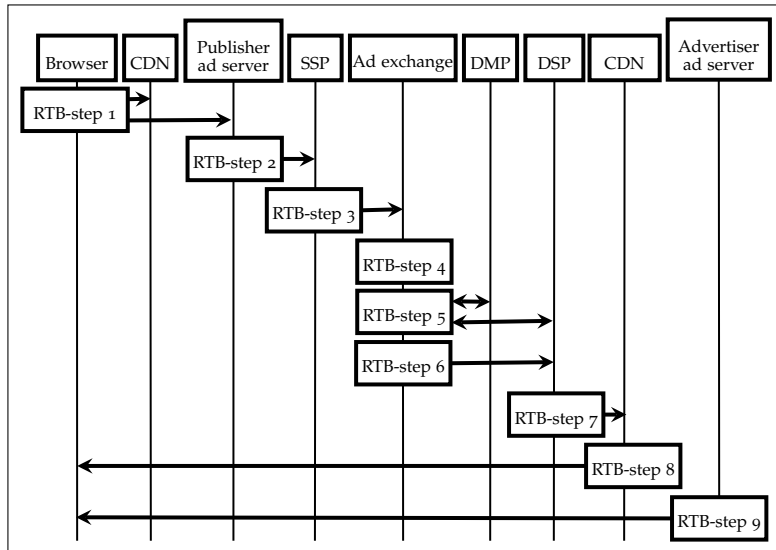


Figure 4.3: RTB sequence diagram with nine RTB-steps.

PROGRESS OF DISCUSSION: AD EXCHANGE

We did not include the term ad exchange (RTB-step 4 - RTB-step 6) as a separate building block in our theoretical view of RTB sys-

²⁹¹ We remark that the RTB-steps above are consistent with the process visualized by Ryan [2018, pp. 6–8] and CM Summit and BattelleMedia [2013]. See also Google [2016a].

²⁹² URL: <http://www.digitaltradingawards.com/digital-trading-awards-usa-2017/best-overall-technology-for-programmatic-trading/rubicon-project-for-best-overall-technology-submission> (29 September 2017).

tems (see Figure 4.2).²⁹³ Instead, the term is included in the RTB sequence diagram (Figure 4.3). The term ad exchange represents a notion with many connotations. An RTB exchange refers to a variety of specialized companies, e. g., RTB auctioneers, mediators, aggregators, and traffic filters (cf. IABureau [2017b, p. 7]).²⁹⁴ It is clear though, that the ad exchange is situated on the supply side (see Subsection 4.3.3).

4.3.10 Section conclusions

From the above, we may draw two section conclusions. First, strategic partnerships are square in the center of RTB systems. The partnerships enable the *real-time integration* of RTB systems. Second, the interconnection of RTB systems and their partners with standardized RTB protocols creates the *infrastructure* for RTB.

More fundamentally, I would like to argue that digital advertising really seems to move toward real-time DMPs as the dominant ad-delivery model (see Subsection 2.3.2.).²⁹⁵ RTB has become more complex and less transparent in terms of visibility to the outside world, as these consist of many closed RTB systems.

Now that we presented a *theoretical* model of RTB systems, we turn to a more detailed *empirical* view of the RTB actors and their interrelationships (Section 4.4). The key objective in that section is to compile an empirical view of RTB systems.

²⁹³ Supra n. 25. We remark that the term 'ad exchange' relates to the concept of 'ad network providers', a term which was frequently used in discussions on the privacy component of Online Behavioral Advertising (OBA). See, e. g., Article 29 Working Party [2010, WP 171, p. 5]: „Behavioural advertising involves the following roles: (a) advertising networks providers (also referred to as "ad network providers"), the most important distributors of behavioural advertising since they connect publishers with advertisers; (b) advertisers who want to promote a product or service to a specific audience; and (c) publishers who are the website owners looking for revenues by selling space to display ads on their website(s)".

²⁹⁴ The full quotation is as follows: „The term 'Exchange' refers to various types of supply intermediaries (e. g., Auctioneers, Mediators, Aggregators, Traffic filters, etc."

²⁹⁵ I am indebted to Van der Hout [2015], Wainberg [2015], and Zorbas [2015] who confirmed this market insight independently from one another. They were, of course, considering this as experts, in their personal capacity.

4.4 EMPIRICAL VIEW OF RTB SYSTEMS

Notwithstanding our remarks on the three parts of our approach - data collection (Part 1, Section 3.1), data reduction (Part 2, Section 3.2), and data modeling (Part 3, Section 3.3), we briefly highlight the three parts taken to collect, reduce, and model the research data for our empirical view.

The course of the section is as follows. In Subsection 4.4.1 we discuss digital media as a contextual data source. In Subsection 4.4.2 we initiate a stateless deep crawl: EU Feeds. In Subsection 4.4.3 we design a construction of our graph. Then, we discuss three empirical observations (Subsection 4.4.4). We end this section with two section conclusions and link the results to the theoretical view (Subsection 4.4.5).

4.4.1 *Digital media as a contextual data source*

We turned to EU Feeds [European Journalism Centre, 2009] as a data source.²⁹⁶ An insight gained from analyzing our first crawls (see Table 3.2) is that data which is collected by a ranking of websites from the Alexa or Quantcast datasets, leads to poor results when attempting to categorize for RTB systems.

In support of our analysis, we remark that Budak et al. [2014] also questioned the view of a by-and-large ad-supported Web. They reported that two-thirds of internet traffic comes from websites that do not show third-party ads. In fact, only 12% of the first parties in their dataset show targeted ads based on information that end-users did not directly provide to the first-party [Budak et al., 2014, p. 13].²⁹⁷

To understand the dependency in the context of online advertising, we take note of the observation by Oremus [2017], a senior technology writer. He illustrated the dependency of news websites relying on the online-advertising business model on his blog as follows.

²⁹⁶ URL: <https://web.archive.org/web/20160908121211/http://www.eufeds.eu/> (29 August 2016).

²⁹⁷ The study has been conducted by analyzing the page views of 13.6 million end-users for the 12 months between June 1, 2013 and May 31, 2014. The data was collected via the Bing Toolbar. Each toolbar installation by an end-user was assigned a UID.

„Print media has been in decline for more than 15 years, its business model obsolesced by the ubiquity of free online content and the rise of online advertising. But all was not lost: The internet brought with it exciting new opportunities and forms. (...) As newspapers withered, digital media ventures - first Slate and Salon, then the Huffington Post, Gawker, Business Insider, BuzzFeed, Mashable, Vice, Vox, Fusion, and countless others - bloomed.” [Oremus, 2017] (emphasis omitted)

Taking the above into account, I assume that there is a need for carefully selecting a context with many advertisements. Therefore, I started to investigate the digital-media context (i. e., a dataset of national and regional European news websites). We reiterate that Van Eijk [2011b] used the same data source to execute *stateful shallow* crawls (see Definition 3.1).²⁹⁸

In the meantime, the assumption of the prevalence of advertising cookies had been investigated by Trevisan et al. [2017] and more recently by Turcios Rodríguez [2018]. Trevisan et al. [2017, p. 8] found that websites categorized as ‘News and Media’ rely more on web tracking than other categories. This was also confirmed by Turcios Rodríguez [2018, pp. 104–109] who found through negative binomial regression that websites categorized by the IABureau as ‘News/Weather/Information’ contain the highest number of third-party trackers that set cookies in the browser.

Having justified the context of our data source, we are now ready to discuss the intricacies of our data-flow collection for our empirical view on RTB.

4.4.2 *Stateless deep crawl: EU Feeds*

The first RTB-step (see Subsection 4.3.9) starts with an HTTP request. We aim to trigger this step by crawling selected webpages with a *stateless deep* crawl and by retaining the research data. The trigger enables us to follow the eight successive RTB-steps.

On 29 August 2016, I submitted a crawl (Crawl17) to the Netograph Cortesi [2017] experimental framework.²⁹⁹ Both the raw Mitmproxy data files (*.mitm) and Netograph’s metadata (*.json)

²⁹⁸ See our definitions for shallow crawl (Definition 3.1), deep crawl (Definition 3.2), stateless web tracking (Definition 2.7), and stateful web tracking (Definition 2.8).

²⁹⁹ For Crawl1, Crawl12, (...) Crawl16 see Table 3.2.

were retained.³⁰⁰ Crawl7 consists of page visits to 8,473 news items from 461 European news websites.³⁰¹ Obviously, the number of newspapers varies per country. United Kingdom is leading with 665 items which were requested from 35 newspapers. Next is Spain (565/32) followed by Italy (548/30), Germany (533/29), and Denmark (475/26).³⁰²

Below, we briefly discuss three intricacies of our *stateless deep* crawl: (A) stateless crawl, (B) EU Feeds as a data source for a deep crawl, and (C) IP address.

A: STATELESS CRAWL

The deep crawl was stateless (see Definition 2.7), so each news item was captured as if an end-user would have visited the webpage for the first time. No browser state was kept between two page visits.

We remark that a stateless crawl suffices in compiling an empirical view of RTB systems because of our focus on the *data-collection* component of privacy (see Section 1.2). Each time a news article is visited the browser appears as a new end-user to the RTB system. In contrast, e. g., for profiling where the focus is more on the *data-application* component of privacy we would recommend a stateful crawl taking into account end-user impersonation.³⁰³

B: EU FEEDS AS A DATA SOURCE FOR A DEEP CRAWL

EU Feeds is an aggregation-web service based on Rich Site Summary (RSS).³⁰⁴ The service collects news articles from the most popular (1) national, (2) regional, and (3) local newspapers in 28

³⁰⁰ The format of the JSON metadata is documented on the Netograph website. URL: <https://netograph.io/docs/formats/details> (1 September 2018).

³⁰¹ Data from two newspapers (38 news items) from Cyprus is not included in our dataset, i. e., Famagusta Gazette, URL: <http://www.famagusta-gazette.com/> (29 August 2016) and Phileleftheros, URL: <http://www.phileleftheros.com/> (29 August 2016).

³⁰² In Luxembourg I crawled 19 news items from one newspaper: L'Essentiel, URL: www.lessestiel.lu (29 August 2016).

³⁰³ See Subsection 3.1.4.

³⁰⁴ European Journalism Centre [2009]: „This source of press articles allows users to get an instantaneous and comprehensive review of the most prevalent issues discussed in each EU Member State, as well as a point of reference for differences in coverage, tone, and outlook surrounding matters of common European concern, as manifested by the media of each country.”

EU countries and is updated every 20 minutes.³⁰⁵ EU Feeds allowed us to collect a snapshot of news items discussed in each country (see Definition 2.7). RSS is a technology allowing a publisher to provide easy access to their readers about new content. RSS allows us to perform a *deep* crawl (see Definition 3.2) of the latest online-news articles.

C: IP ADDRESS

I remark that I did not use OpenVPN [Yonan, 2001] endpoints which would have enabled us to crawl a webpage with an IP address matching the webpage's country origin.³⁰⁶ Instead, I crawled the news websites with a headless browser originating from Amazon AWS IP addresses. Furthermore, I remark that Trevisan et al. [2017] visited websites from nine EU countries,³⁰⁷ and found that the number of third-party cookies does not change. Furthermore, we remark that Turcios Rodríguez [2018] visited websites from 15 EU countries,³⁰⁸ and found that websites seem to follow the local law with respect to setting a cookie in a browser, even when a website is visited from different countries with OpenVPN endpoints.

4.4.3 Construction of the graph

Obviously, graphs to understand RTB systems can be constructed in many ways. I briefly highlighted ten different approaches to WPM visualizations in Section 2.7. Our approach to analyze Crawl7 data is through the lens of the generic web-tracking model (see Figure 3.2). I stored the network of RTB partners in a separate graph database and did so by country. Below, we briefly discuss two intricacies (called M-steps) of our approach to model the graph from data-flow collection to small-data modeling: (A) constructing the referrer graph, and (B) graph refactoring.

³⁰⁵ The list of EU Feeds-countries includes Norway (not a member of the EU) and all 28 EU member states except Croatia.

³⁰⁶ Supra n. 224.

³⁰⁷ They are: France (FR), Italy (IT), Germany (DE), Finland (FI), Netherlands (NL), Portugal (PT), Spain (ES) and Sweden (SE).

³⁰⁸ They are: Austria (AT), Belgium (BE), Czech Republic (CZ), France (FR), Germany (DE), Greece (GR), Hungary (HU), Italy (IT), Netherlands (NL), Poland (PL), Portugal (PT), Romania (RO), Spain (ES), Sweden (SE), and United Kingdom (UK).

M-STEP 1: CONSTRUCTING THE REFERRER GRAPH

The construction of the graph is based on the HTTP headers retained as Netograph's metadata files (*.json).³⁰⁹ The meaning of the edges is to highlight a relation between two partners (the nodes in the graph) in an RTB network. The importance of the (distinct-directed) edges is to express valuable information about the data flow. Similarly, the distinct nodes provide us with information about the companies processing the tracking data.

The following four types of HTTP metadata were used to construct the referrer graph.

- (1) Host of the resource.
- (2) Referrer-request header.
- (3) Response-code response header.
- (4) Location-response header.

Below we discuss the role of these four HTTP-header fields in relation to the construction of the graph (denoted by number*).

The first two fields define the direction of an edge. If the HTTP header contains (1*) a host of the resource (V_A) and (2*) a referrer-request header (V_B), then the edge (E) of a directed subgraph denoted by $g = (V, E)$ is retained as $E: V_A \leftarrow V_B$ (i. e., V_B refers to V_A).

The third and fourth field provide us with information about HTTP redirection (Subsection 3.3.5:B1). For (3*) it holds: if no referrer-request header is present, we look for a response-code response header (third field) and we extract the corresponding location-response header (fourth field). The edge (E) is then retained as $E: V_A \rightarrow V_C$ (i. e., V_A redirects to V_C). For (4*) it holds: if no location-response header is found (which happens occasionally), we retain the edge (E) as $E: V_A \rightarrow V_A$ (i. e., a *self-referencing loop*).

M-STEP 2: GRAPH REFACTORING

DEFINITION 4.13: GRAPH REFACTORING is defined as an activity of merging the origins of the webpages that are visited in a crawl onto a single node in a graph.

³⁰⁹ Although the Netograph framework is still in alpha, the documentation for the JSON format - the metadata format of our research data - can be found online. URL: <https://netograph.io/docs/formats/details> (29 January 2018).

The aim of refactoring our graph is twofold: (1) to simplify its structure and (2) to focus on the relationships between the resources present on a webpage. To reach this goal I replaced the domain name of the webpage containing a news item with the label 'crawled.io'. In other words, refactoring a graph means *folding* the graph by placing the crawled webpages onto one single node. The importance of a refactored graph is that it enables a comparison of the research data (Crawl7) across countries.

PROGRESS OF DISCUSSION: EXPRESSING ADDITIONAL RELATIONSHIPS

We already noted at the beginning of this subsection that there are many ways in which the relationships can be constructed between nodes in a graph. Therefore, we should also look at the question: how can we improve the referrer model? To guide our discussion we will refer to Listing 4.2.

Listing 4.2: Example of (improved) Netograph metadata.

```

1  (...)
2  "browser_initiator": {
3    "linenumber": 119,
4    "type": "parser",
5    "url": "https://www.natuurlijkehaarkleuring.nl/afspraak/"
6  },
7  "browser_type": "Script",
8  "document_url": "https://www.natuurlijkehaarkleuring.nl/
   afspraak/",
9  "host": "d3gxy7nm8y4yjr.cloudfront.net",
10 "request": {
11   "headers": {
12     "Accept": "*/*",
13     "Accept-Encoding": "gzip, deflate",
14     "Connection": "keep-alive",
15     "Host": "d3gxy7nm8y4yjr.cloudfront.net",
16     "User-Agent": "Mozilla/5.0 (X11; Linux x86_64) AppleWebKit
   /537.36 (KHTML, like Gecko) HeadlessChrome
   /63.0.3239.132 Safari/537.36"
17   },
18   "method": "GET",
19   "referrer_policy": "no-referrer",
20   "url": "https://d3gxy7nm8y4yjr.cloudfront.net/js/embed.js"
21 },
22 (...)
```

Nowadays, website resources are often served from a Content Delivery Network (CDN). Listing 4.2 contains an example of such a case. The example shows a JavaScript named ‘embed.js’ (Listing 4.2, r. 7 and r. 20) served from a CloudFront CDN (Listing 4.2, r. 9). We note that the security header ‘referrer_policy’ is set which means that no referrer header (Listing 4.2, r. 22) is present. Serving content this way implies that the *host* field (Listing 4.2, r. 9) *may* not always be the same as the *document.url* (Listing 4.2, r. 8) of the resource.

Here, we remark that Hearne [2013] was (one of) the first scholars to take the relationship between the host (V_A) and the *document.url* (V_D) into account. We believe that our referrer model may benefit from the expression of an additional relationship (i. e., an edge) between the two nodes. The edge (E) is then retained as $E: V_A \rightarrow V_D$ (i. e., V_A *embeds* V_D). The importance of the relationship is the fact that it highlights the presence of the JavaScript ‘embed.js’ in a first-party context (i. e., Listing 4.2, r. 8).

Recently Bashir and Wilson [2018] also considered the relationship between DOM elements embedded in a first-party context (e. g., (1) an image tag or (2) an iFrame injected by a JavaScript in a first-party context). They proposed a different type of graph, i. e., the *inclusion graph* [Bashir & Wilson, 2018, p. 90]. The meaning of the edges in the inclusion graph is tailored to the expression concerning the relationships between all elements in their *own* context, i. e., not the *inferred* context induced by the metadata in a *referrer graph*. For instance, a JavaScript’s origin V_E injected in the browser DOM within the context of an iFrame V_F is expressed as an edge (E) and retained as $E: V_E \leftarrow V_F$ (i. e., V_F *embeds* V_E).

Bashir and Wilson [2018, p. 100] concluded that a referrer graph fails to capture the relationship between DOM elements injected in a first-party context. My opinion is different. (1) I fully agree that the proposed inclusion graph differs from a graph based on just referrer metadata. However, as we have discussed in this subsection, (2) the representation of *redirection metadata* in combination with (3) *graph refactoring* helps us to improve our understanding of web tracking (e. g., RTB systems and RTB partner networks). Hence, with the two new means (redirection metadata and graph refactoring) we are still able to capture the relationship between DOM elements injected in a first-party context.

PROGRESS OF DEVELOPMENT: NETOGRAPH METADATA

At this point, we remark that Netograph has shifted toward using the Chrome debugging protocol (Subsection 3.1.2:C) instead of Mitmproxy (which we used at the time of collecting the data for Crawl7). A direct consequence is that the metadata files that come with the improved Netograph experimental framework, now contains the precise location of the HTML code (Listing 4.2, r. 3). This is the code that triggered the loading of a web resource 'embed.js' (Listing 4.2, r. 20) in the first party context (Listing 4.2, r. 5). The source of the metadata is the (improved) Chrome DevTools network analysis feature [Basques, 2018]. With this metadata we will be able to express additional relationships in WPM graphs. The expression includes, e. g., (1) JavaScript cookies, (2) JavaScript pixels, (3) websocket connections, and (4) iFrames injected into the browser DOM by RTB network partners.³¹⁰

The progress of development gives us a clear view of the steps taken in the process from data collection to data modeling (Subsection 4.4.1 – Subsection 4.4.3). Below we continue our empirical view of RTB systems with three observations.

4.4.4 *Three empirical observations*

An important (next) step (in data science) is to familiarize ourselves and the end-users with the dataset. Therefore, we report three observations below.

The first observation is a dendrogram depicting (1) the similarities between the top-20 edges and (2) the number of occurrences (in the HTTP header) of an edge in a country by a heatmap. A distinct edge is an expression of a clear relationship in the referrer model (see the previous Subsection 4.4.3). The number of occurrences of the edge is added as a weight to the edge. The second observation is the list of the prevalence of eight leading companies present on European news websites. The third observation is an example of cross-border web tracking by zooming in on the EU presence of the Rubicon Project. After the report of the three observations, we present two section conclusions (Subsection 4.4.5).

³¹⁰ See also Figure 4.1 (i. e., image, iFrame, script).

OBSERVATION 1: SIMILARITIES BETWEEN THE TOP-20 EDGES

The '20x27'-rectangle (top-20 edges x the 27 countries) in Figure 4.4 (next page) is visualized as (1) a *heatmap* and (2) a *dendrogram*.³¹¹ I admit that the combined visualization may appear complex at first sight. To guide our observation on the similarities between the top-20 edges, we will partition the complexity into: (A) a description of the '20x27'-rectangle, (B) an explanation of the heatmap, and (C) a guidance for reading the dendrogram.

A: '20x27'-RECTANGLE

On the *x-axis* there are 20 distinct edges (a *subset* of the total number of 13,789 distinct edges in Crawl7) denoted by number* (the star denotes that the number is an edge). On the *y-axis* we see the countries listed in alphabetical order. The abbreviation for each country is denoted by the language code (see also Table 4.3a, p. 168).³¹² The relation between an edge and a country is given by color. White means: no presence in that country. Blue means: presence in that country. The intensity of blue is related to the number of occurrences in that country (the more intensive, the higher the number).

We note that the dataframe containing all distinct edges from Crawl7 was sorted by the total count of the presence of a distinct edge in a country. For instance, the edge in the last column (Edge 20*) is present in all countries (see Figure 4.4).

As we can now read, the subset (called *top-20 edges*) contains six edges with a presence in 27 countries, eight edges (with a presence in 26 countries), and six edges (with a presence in 25 countries). These numbers correspond with the number of empty (white) cells in the rectangle (see Figure 4.4). We will list each top-20 edge in our discussion on the dendrogram (C, pp. 165–166). For now, we remark that the top-20 edges belong to the three leading companies (1) Google, (2) Twitter, and (3) Facebook.

³¹¹ Figure 4.4 was created with the R-module Heatmaply [Galili, Tal, O'Callaghan, Alan, Sidi, Jonathan, Sievert, & Carson, 2017]. URL: <https://github.com/talgali/heatmaply> (4 September 2018).

³¹² We assigned a well-known two-letter country code as an abbreviation for the use of Member states of the EU. See, e.g., URL: http://ec.europa.eu/eurostat/statistics-explained/index.php/Glossary:Country_codes (16 March 2018). For courtesy, we provide the country names in full in brackets after the abbreviation in Table 4.3a.

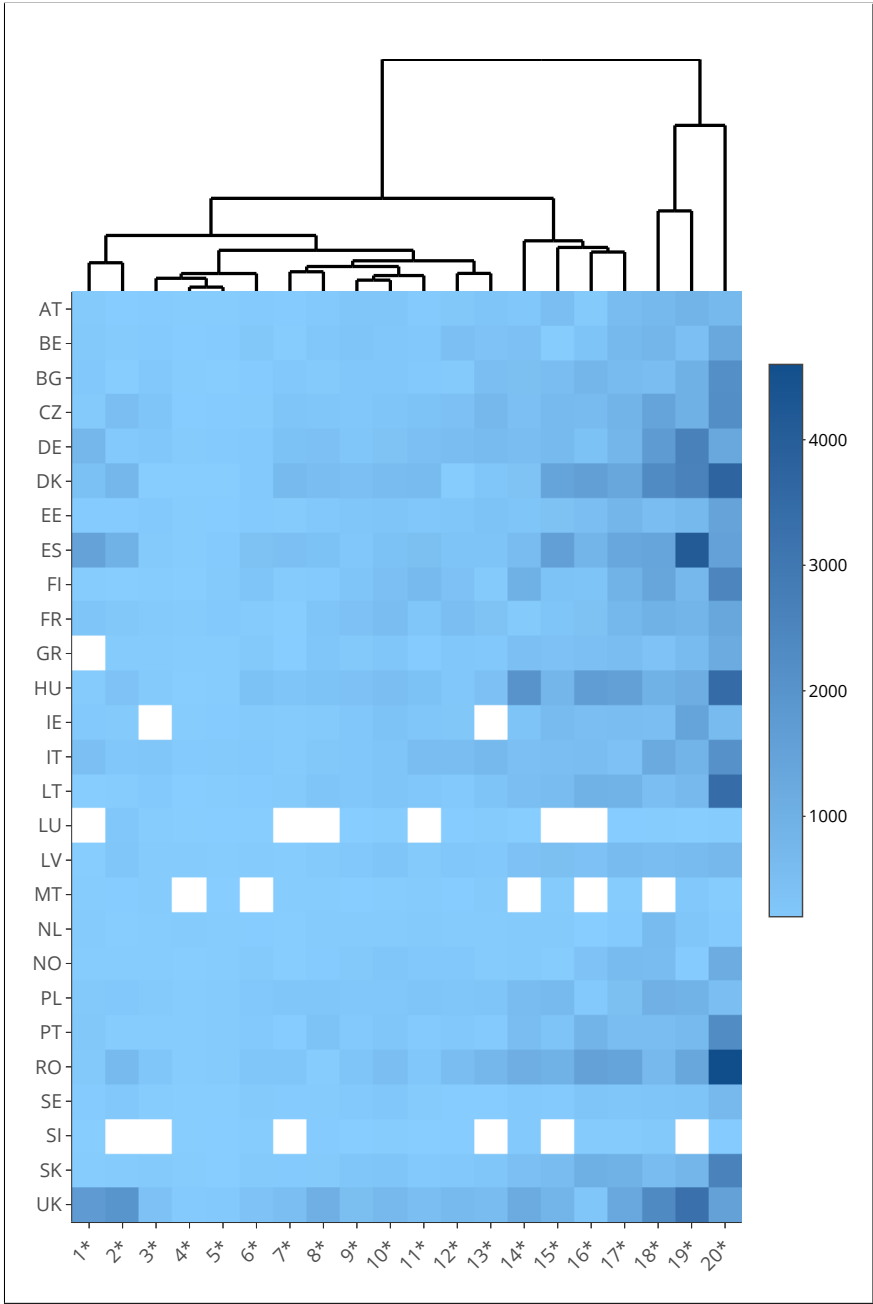


Figure 4.4: EU heatmap and dendrogram.

In Observation 2 (prevalence of eight leading companies) we will refer to a more elaborate subset of edges, a total of 100 distinct edges from Crawl7 (called *top-100 edges*). The subset top-100 edges contains six edges with a presence in 27 countries, eight edges (in 26 countries), 15 edges (in 25 countries), 25 edges (in 24 countries), 15 edges (in 23 countries), 11 edges (in 22 countries), 18 edges (in 21 countries), and two edges (in 23 countries).

B: HEATMAP

A heatmap enables us to visually compare the number of occurrences of a distinct edge per country. This means that the number of occurrences in the graph model of the edges in a country is represented by a color shade. A high(er) number corresponds with a dark(er) color. Likewise, low numbers of a distinct edge within a specific country result in a light(er) color. The intensity bar on the right of the visualization serves as an indication of the number of distinct edges. For instance, the cell in the top-right corner (Edge 20*, 'AT') represents a total of 751 occurrences in the graph model which corresponds with a light intensity of the color. In contrast, the darkest cell in the same column (Edge 20*, 'RO') represents a total of 4,626 occurrences.³¹³ The reader is invited to check the variations in occurrences in the heatmap himself.³¹⁴

C: DENDROGRAM

A dendrogram is a graphical representation of the relationships of similarities among a group of entities. In our case, the dendrogram enables us to visually compare the presence of a distinct edge across countries by the number of occurrences. For instance, the three columns on the right end of the x-axis (Edge 18*, 19*, and 20*) group together in a cluster of three distinct edges due to their similarity in presence across the 27 European countries.

We discussed the characteristics of the edges in our graph in M-step 1, constructing the referrer graph (Subsection 4.4.3). We recall that the meaning of the edges is to highlight a relation between

³¹³ The edge with the highest number of occurrences (a total of 5,703) is eas3.emediate.se -> crawled.io. Please note that Emediate is „the leading provider of ad serving technology in the Nordic region.“ Clearly, due to this limited country presence it is not shown in our heatmap. URL: <https://eas3.emediate.se/> (15 September 2018).

³¹⁴ See n. 36.

two partners (the nodes in the graph) in an RTB network. In other words, the importance of the distinct edges is to express valuable information about the data flow *through* the RTB network as RTB partners refer or redirect to each other.

On top of the heatmap there is visualized a tree (a *dendrogram*). This tree groups all countries and (importantly!) shows the difference among the relations of the groups. There are 19 clusters of edges (see Figure 4.5), viz. one cluster of 20 edges, one cluster of 17 edges, one cluster of 13 edges, one cluster of 11 edges, one cluster of 7 edges, one cluster of 5 edges, two clusters of four edges, four clusters of three edges, and seven clusters of two edges.

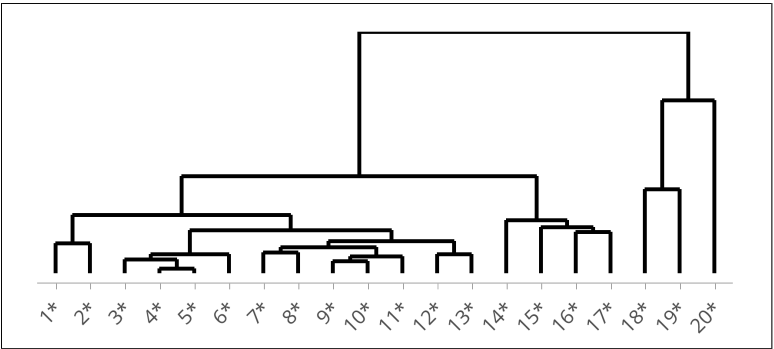


Figure 4.5: EU dendrogram.

Two clusters are important: the cluster of 17 edges and the cluster of three (one of the four, in our case, the rightmost cluster). The height of the dendrogram lines (i. e., leaves) indicate the degree of difference between the distinct edges. The significance of the height of the branches is explained by Wheaton Lexomics [2012] as follows.

„The arrangement of the clades tells us which leaves are most similar to each other. The height of the branch points indicates how similar or different they are from each other: the greater the height, the greater the difference.” [Wheaton Lexomics, 2012, 3’:05"-3’:22"]

We remark that a *clade* is a group of elements that consists of a common ancestor and all its lineal descendants. It represents a

single 'branche' in the dendrogram. It is easy to identify a clade using a phylogenetic tree. Furthermore, we remark that the order of the leaves is irrelevant in a dendrogram and does not provide us with additional information (cf. Wheaton Lexomics [2012]).

The relationships of similarities between the top-20 edges provides us information about the last three edges (18*-20*, henceforth Chunk A*):

- (18*) `www.google.analytics.com -> crawled.io`,
- (19*) `pagead2.googlesyndication.com -> crawled.io`,
- (20*) `static.xx.fbcdn.net -> www.facebook.com`.

The edges in Chunk A* are very similar and chunk together as a first group in the branching diagram. Furthermore, the distribution of the edges in Chunk A* is substantially different from the distribution in the remaining chunks.

Similarly, the dendrogram provides us information about the next tallest group of leaves (14*-17*, henceforth Chunk B*).

- (14*) `fonts.gstatic.com -> fonts.googleapis.com`,
- (15*) `googleads.g.doubleclick.net -> crawled.io`,
- (16*) `scontent.xx.fbcdn.net -> www.facebook.com`,
- (17*) `www.facebook.com -> crawled.io`.

The edges in Chunk B* are very similar and chunk together as a second group in the branching diagram. The distribution in Chunk B* is different from the distribution in Chunk A*.

Moreover, we turn our attention to the group of the first two edges (1*, 2*, henceforth Chunk C*):

- (1*) `tpc.googlesyndication.com -> crawled.io`,
- (2*) `s0.2mdn.net -> s0.2mdn.net`.

The edges in Chunk C* (the edges denoted by 1* and 2*) are also very similar. The reader is invited to check the leaves in the dendrogram himself.

Here, we remark again - taking into account the measurement against the height of chunks A*, B*, and C* - that the distribution of the edges of these three chunks are substantially different from the distribution in the remaining chunks. Therefore, based on the dendrogram we may conclude that the nine edges (1*-2*, 14*-17*, and 18*-20*) in chunks C*, B*, and A* can be viewed as the top nine edges in both number of occurrences and similarities.

For completeness, we identified the remaining 11 top-20 edges (3*-13*) present in our dataset. We mention them in one group below and call them Chunk D*.

- (3*) apis.google.com -> accounts.google.com,
- (4*) platform.twitter.com -> platform.twitter.com,
- (5*) syndication.twitter.com -> crawled.io,
- (6*) fonts.googleapis.com -> crawled.io,
- (7*) tpc.googlesyndication.com -> tpc.googlesyndication.com,
- (8*) www.google.com -> crawled.io,
- (9*) connect.facebook.net -> crawled.io,
- (10*) staticxx.facebook.com -> crawled.io,
- (11*) stats.g.doubleclick.net -> crawled.io,
- (12*) platform.twitter.com -> crawled.io,
- (13*) apis.google.com -> crawled.io.

AN EXAMPLE OF A TOP-20 EDGE: DOUBLECLICK

As a rather arbitrary example we take edge 2*. The edge (2*) 's0.2mdn.net -> s0.2mdn.net' represents a Google DoubleClick edge referring to a resource on the same host. The self-referencing edge is present in 26 countries (not in Slovenia, see the white colored block) in the '20x27'-rectangle. In fact, the DoubleClick self-referencing edge in this example is a mini-web page.³¹⁵

The size of the mini-web page is 300 pixels wide by 250 pixels tall, better known as a medium rectangle ad-slot (see Table 4.1). The RTB process is initiated by a JavaScript embedded in the mini-web page.³¹⁶ The relevant information now is that the HTTP-referrer field of the HTTP-header (GET request) of the JavaScript points to the origin of the medium rectangle ad-slot. So, the (full) relationship captured in the graph model is well known (i. e., $E: V_{s0.2mdn.net} \rightarrow V_{s0.2mdn.net}$).

OBSERVATION 2: PREVALENCE OF EIGHT LEADING COMPANIES

WPM scholars often include a (brief) report on the prevalence of the leading companies in their dataset. Three examples are Yu et

³¹⁵ Media Innovation Group's mini-web page: URL: <https://s0.2mdn.net/3722876/1471541890008/BA-SM1B-Hands-300x250/300x250.html> (2 February 2018)

³¹⁶ URL: <https://s0.2mdn.net/3722876/1471541890008/BA-SM1B-Hands-300x250/300x250.js> (2 February 2018).

al. [2016, p.5], Libert, Graves, and Nielsen [2018, pp. 3–5], and Turcios Rodríguez [2018, p. 75]. I take a (slightly) different approach that enables the reader to perform a manual lookup. Below, I report on the prevalence of the leading companies in Crawl7. The *nodes* of our small-data model represent the Fully Qualified Domain Name (FQDN).³¹⁷ This FQDN enables us to perform a manual lookup of the companies belonging to the distinct nodes of the top-100 edges.

We reiterate that the companies (1) Google, (2) Twitter, and (3) Facebook are present in at least 25 European countries (Figure 4.4). We expand our view to the companies belonging to the FQDN of the nodes of the distinct *top-100 edges*. Hence, we report a company prevalence in at least 20 European countries of these three organizations.

The eight leading companies (denoted by number*) in our top-100 edges (200 nodes) are:

- (1*) Google, with 91 distinct nodes, followed by
- (2*) Twitter (14 nodes),
- (3*) Facebook (12 nodes),
- (4*) the Rubicon Project (7 nodes);
- (5*) Crownpeak [Evidon] (4 nodes),³¹⁸
- (6*) Oracle [AddThis] (3 nodes),³¹⁹
- (7*) Turn (3 nodes), and
- (8*) Yahoo [BrightRoll] (3 nodes).³²⁰

Google's presence on the first place in the data may not come as a surprise.³²¹ Digital media may combine the delivery of videos, advertising, and analytics with, e. g., YouTube, Google Analytics, and DoubleClick (1*).

Obviously, social media components are numerous, e. g., event tracking by Twitter's tweet-button (2*), Facebook's like-button (3*), or Oracle's addThis-button (6*). Based on the context of our crawl, we categorize Crownpeak (5*) as data relating to their consent services which is based on notice and choice for cookies [Meyer, 2017a]. Furthermore, we categorize the Rubicon Project (4*), Turn (7*), and Yahoo (8*) as RTB data.

³¹⁷ For instance, the FQDN 'o.2mdn.net' (which belongs to Google [DoubleClick]) specifies the exact location of the hostname 'o' in the parent domain '2mdn.net'.

³¹⁸ Crownpeak acquired Evidon recently [Meyer, 2017b].

³¹⁹ Oracle acquired AddThis in 2016 [Tawakol, 2016].

³²⁰ Yahoo acquired BrightRoll in 2014 [Meron & Huh, 2014].

³²¹ See note 257

OBSERVATION 3: EU PRESENCE OF THE RUBICON PROJECT

Recently, the prevalence of leading companies in different countries has caught the attention of WPM scholars (see, e.g., Iordanou et al. [2018], who reported on *cross-border web tracking*). For Observation 3, we provide an example of cross-border web tracking by zooming in on one of the eight leading companies, i.e., the Rubicon Project (see Table 4.3).

Table 4.3: EU presence of the Rubicon Project.

COUNTRY	OCCURRENCES		
UK (United Kingdom)	18,741		
CZ (Czech Republic)	14,647		
FR (France)	12,543		
BE (Belgium)	9,757		
ES (Spain)	8,927		
HU (Hungary)	7,783		
FI (Finland)	7,717		
DK (Denmark)	7,487		
RO (Romania)	4,100		
LT (Lithuania)	4,041	TYPES	TOTAL
PT (Portugal)	3,753	Min.:	22
EE (Estonia)	1,908	1st Qu.:	369
IT (Italy)	1,687	Median:	1,687
AT (Austria)	1,468	Mean:	4,355
NO (Norway)	1,397	3rd Qu.:	7,717
IE (Ireland)	614	Max.:	18,741
SE (Sweden)	564	Total:	108,875
PL (Poland)	454	Distinct edges:	252
DE (Germany)	369	Distinct nodes:	273
LV (Latvia)	366		
SK (Slovakia)	228		
GR (Greece)	216		
BG (Bulgaria)	46		
NL (Netherlands)	40		
MT (Malta)	22		
LU (Luxembourg)	0		
SI (Slovenia)	0		

(a)

(b)

Crunchbase [2007], a website with information about business describes the company as follows: „the Rubicon Project is an advertising automation platform enabling premium publishers to transact advertising brands.”

Table 4.3 consists of two tables (Table 4.3a and Table 4.3b). Table 4.3a shows the 27 European countries in Crawl7.³²² In the first column ‘Country’, the abbreviation for each country is denoted by the language code.³²³ The second column ‘Occurrences’ shows the total number of occurrences of the distinct edges related to the Rubicon Project in a country. We have ordered the list according to the occurrences. In Table 4.3b we provide a numerical summary with nine types of measures: their meaning is in the first column ‘Types’ and the corresponding result in the second column ‘Total’.

The prevalence in European countries of the Rubicon Project follows from Table 4.3a. First, we remark that the number of distinct edges (252) and distinct nodes (273) in the second column ‘Total’ in Table 4.3b are in fact two indicators for the close relationship between the Rubicon Project’s partners. The number of distinct edges and distinct nodes denote the interconnectedness of partners within the Rubicon Project partner network.

Second, the Rubicon Project is omnipresent in 25 European countries (Table 4.3a, rr. 1–25). The Rubicon Project is leading in the United Kingdom (UK) with 18,741 occurrences, followed by the Czech Republic (CZ) (14,647), and France (FR) (12,543). We encountered only two countries in Crawl7 without a Rubicon Project presence, i. e., Luxembourg (LU) and Slovenia (SI) with zero occurrences (Table 4.3a, rr. 26–27).³²⁴

PROGRESS OF DISCUSSION

I presented a *heatmap* of the top-20 edges at the Dagstuhl seminar 17162, Online Privacy and Web Transparency [Van Eijk, 2017]. In principle, a heatmap in itself is similar to the dendrogram (see Figure 4.4). However, without a branching diagram that represents the relationships of similarity among a group of entities it was difficult to transfer the knowledge. The insight gained from the workshop was that the relationship between the edges could ben-

³²² See n. 305.

³²³ See n. 312.

³²⁴ We remark that the absence of a presence by the Rubicon Project in Luxembourg (LU) and Slovenia (SI) is not visible in Figure 4.4 (which depicts the ‘top-20 edges’).

efit from further clarification. This led to the *dendrogram* depicting the edge similarities (see Figure 4.4).

4.4.5 Section conclusions

When it comes to what we now have learned with respect to RQ2, we start our explanation at Figure 4.2. That picture represents what is known. We immediately admit that this is theory. Indeed, it is not yet precisely known how RTB systems work in practice. This is the precise reason why publishers when monetizing their news items with RTB ads are not able - in our view - to provide end-users with information that is sufficiently specific for a valid consent.

Our two section conclusions are based on:

- (1) the total number of HTTP requests (occurrences) to network partners on European news websites; this total number outnumbers the total number of requests to non-advertising related web resources; the reason is the interrelationships of the network partners, and
- (2) the relationships of similarities between distinct edges.

SECTION CONCLUSION 1

Our first section conclusion is that the network of partners *automatically emerges from the data*, when analyzed through the lens of a refactored referrer graph. Here, we reiterate that we are interested in a network of partners (see Section 4.1) who collectively form an RTB system (see Definition 4.2).

SECTION CONCLUSION 2

Our second section conclusion is that the leading RTB companies are omnipresent in most EU countries. This prevalence gives them a strategic *cross-border* position. The capabilities in terms of data collection of the leading RTB are amplified even more if we take (cross-border) partner networks into account.

4.5 CONSENT AS A PRIVACY COMPONENT

Consent is an important privacy component from a WPM perspective when collecting research data. We mention Libert and Graves [2018], Trevisan et al. [2017] as recent attempts and Leenes and Kosta [2015] as an early attempt to investigate consent from a WPM perspective. Furthermore, Libert et al. [2018] recently published a factsheet on the changes in seven EU countries in terms of the number of third-party cookies on European news websites after the GDPR.³²⁵ They reported an average decrease of 22% of third-party cookies per page across all news websites visited in a *stateless shallow* crawl (Definition 3.1).³²⁶ A total of 194 European media websites were visited with the WebXray experimental framework (Subsection 2.6.4) before and after the GDPR entered into force.³²⁷

In this section, we present a *legal* view on consent as a privacy component. It corresponds with Step 7 of our research methodology, viz. compilation a normative framework (see Section 1.5). First, we briefly discuss consent as a legal basis under the GDPR (Subsection 4.5.1). Then, in Subsection 4.5.2 we concisely discuss the changes since the year 2002 in the legal norm with respect to storing and reading a information (cookies and similar technologies). In Subsection 4.5.3, we differentiate between two types of consent: (1) a strict consent mechanism and (2) an implied consent mechanism. Moreover, we analyze the two types of consent. The aim is to identify differences in consent implementations for tracking cookies in European countries. To do so, we rank the countries on a scale in Subsection 4.5.4. In Subsection 4.5.5 we provide section conclusions.

4.5.1 *Consent as a legal basis under the GDPR*

Zuiderveen Borgesius [2015] argued that the legal norm stipulated in Article 5(3) e-Privacy Directive (EPD) does not provide a legal basis for the processing of personal data. Analogously to his

³²⁵ They are (1) Finland (FI), (2) France (FR), (3) Germany (DE), (4) Italy (IT), (5) Poland (PL), (6) Spain (ES), and (7) the United Kingdom (UK) [Libert et al., 2018, p. 2].

³²⁶ See n. 298.

³²⁷ The crawls were initiated in April 2018 and July 2018. The number of websites per EU country is as follows: Finland (20), France (30), Germany (30), Italy (31), Poland (29), Spain (33), and United Kingdom (31) (cf. Libert et al. [2018, p. 6]).

reasoning, we will argue that the legal basis for processing personal data for the purpose of RTB nowadays shall (in most cases) be based on Article 6(a) GDPR.

The GDPR adds - in comparison to the GDPR - a new privacy criterion for consent as a legal ground for processing personal data, i.e., affirmative action by the end-user. The change in the GDPR is relevant to our discussion, since the e-Privacy Regulation (EPR) (see Subsection 4.5.2) is a *lex specialis* to the GDPR (*lex generalis*).³²⁸ Nowadays, the five criteria for valid consent stipulated in Article 4 sub 11 GDPR are:

- (1) free,
- (2) specific,
- (3) informed,
- (4) unambiguous, and
- (5) affirmative action.³²⁹

Furthermore, the Art. 29 WP reiterates in its guidelines on consent under Regulation 2016/679 [Article 29 Working Party, 2018, WP 259 rev.01] that consent should be given *prior* to the processing activity. The full quotation is as follows.

„Although the GDPR does not literally prescribe in Article 4(11) that consent must be given *prior* to the processing activity, this is clearly implied. The heading of Article 6(1) and the wording ‘has given’ in Article 6(1)(a) support this interpretation. It follows logically from Article 6 and Recital 40 that a valid lawful basis must be present before starting a data processing. Therefore, consent should be given prior to the processing activity.” [Article 29 Working Party, 2018, WP 259 rev.01, pp. 17–18] (emphasis added)

4.5.2 Consent under the EPD/EPR

Consent under the EPD overlaps with the concept of consent under the GDPR (see, e.g., Kamara and Kosta [2016, p. 11]). In fact,

³²⁸ See EPR European Commission [2017, p. 2].

³²⁹ Article 4 sub 11 GDPR: „‘consent’ of the data subject means any freely given, specific, informed and unambiguous indication of the data subject’s wishes by which he or she, by a statement or by a clear affirmative action, signifies agreement to the processing of personal data relating to him or her.”

consent for tracking cookies and similar technologies has been - at the time of crawling (August 2016) - a legal requirement under ePrivacy Directive 2002/58/EC [Parliament of the EU and the Council, 2002] amended by 2009/136/EC [Parliament of the EU and the Council, 2009]. Informed consent is required *prior* to storing tracking cookies or reading information from the end-user's browser. Access to the browser is only permitted (1) after an end-user has been provided with clear and comprehensive information and (2) an end-user has subsequently understood the information plus the consequences of what he is consenting to.³³⁰

The Art. 29 WP provided further guidance for cookies in various opinions, e.g., on online behavioral advertising [Article 29 Working Party, 2010, WP 171], on cookie-consent exception [Article 29 Working Party, 2012, WP 194], on cookies [Article 29 Working Party, 2013b, WP 208], on fingerprinting [Article 29 Working Party, 2014c, WP 224], and with the results of a joint cookie sweep by the Data Protection Authorities DPAs [Article 29 Working Party, 2015a, WP 229]. This development leads us to two remarks.

First, we remark that the legal norm for the protection of information stored in and related to an end-user's terminal equipment has shifted from *opt-out* [Parliament of the EU and the Council, 2002, 2002/58/EU],³³¹ via *opt-in* [Parliament of the EU and the Council, 2009, 2009/136/EU], to *prohibited/except* in the European Commission's proposal for the ePR [European Commission, 2017].³³²

³³⁰ Article 5 sub 3 EPD: „Member States shall ensure that the storing of information, or the gaining of access to information already stored, in the terminal equipment of a subscriber or user is only allowed on condition that the subscriber or user concerned has given his or her consent, *having been provided with clear and comprehensive information, in accordance with Directive 95/46/EC, inter alia, about the purposes of the processing.* This shall not prevent any technical storage or access for the sole purpose of carrying out the transmission of a communication over an electronic communications network, or as strictly necessary in order for the provider of an information society service explicitly requested by the subscriber or user to provide the service." (emphasis added)

³³¹ Amendment 128 of the first reading of the European Parliament proposed the modification of Article 5 sub 3 of the EPD (2002/58/EU) as a *prohibition* [Dumortier & Kosta, 2015, p. 55].

³³² Article 8 sub 1 European Commission [2017]: „The use of processing and storage capabilities of terminal equipment and the collection of information from end-users' terminal equipment, including about its software and hardware, other than by the end-user concerned shall be *prohibited*, except on the following grounds:
(a) it is necessary for the sole purpose of carrying out the transmission of an electronic communication over an electronic communications network; or
(b) the end-user has given his or her consent; or

Second, we remark that at the time of writing, the European legislative procedure of the EPR is about to enter the next phase, i. e., the Trilogue negotiations. The purpose of the Trilogue negotiations is to enable the co-legislators (the European Parliament, the Council of the European Union, and the European Commission) to reach agreement on the text of the EPR.³³³ In this phase - as we have witnessed in the Trilogue negotiations of the GDPR - there can still be significant changes to the (final) text.

4.5.3 *Two types of consent implementations*

In 2016, EU publishers implemented two different types of (technical) mechanisms to meet the legal requirement of informed consent for cookies: (A) a strict-consent mechanism, or (B) an implied-consent mechanism. The reasons for the difference between A and B are (mainly) due to subtle differences in transpositions of the ePrivacy Directive into national laws. First of all, we refer to Cofone [2017], Cofone [2015, pp. 181–184] and the comparative analysis produced by the law firms DLA Piper [2016] or Fieldfisher [2014]). Below, we discuss both mechanisms briefly.

A: STRICT-CONSENT MECHANISM

DEFINITION 4.14: A STRICT-CONSENT MECHANISM implements the legal requirements in such a way that no tracking cookies must be placed on the end-user's device or read from it - when he requests a webpage (viz. RTB-step 1).

Therefore, a strict implementation of e. g., a consent banner or a cookie wall would not allow storage of cookies or access to metadata already stored on the end-user's device *prior* to an end-users informed consent. Strict implementations would block us from collecting RTB data as research data.

(c) it is necessary for providing an information society service requested by the end-user; or

(d) if it is necessary for web audience measuring, provided that such measurement is carried out by the provider of the information society service requested by the end-user." (emphasis added)

³³³ See, e. g., URL: <http://www.europarl.europa.eu/ordinary-legislative-procedure/en/interinstitutional-negotiations.html> (5 September 2018).

B: IMPLIED-CONSENT MECHANISM

DEFINITION 4.15: An IMPLIED-CONSENT MECHANISM already stores or reads information from the browser when an end-user requests a webpage although he is given notice with a consent banner.

In contrast to the strict-consent mechanism, implied-consent mechanisms implemented the legal requirements less strictly. Implied-consent mechanisms suggest valid consent as a logically necessary consequence of visiting the webpage. Consent is not directly expressed by, e. g., ticking a box in a consent banner. Instead, the legal validity is assumed to be inherent in the nature of requesting a news item and/or continued browsing by the end-user.

4.5.4 *Analysis of consent implementations*

So far, we identified differences in consent implementations for tracking cookies in European countries. We note that Turcios Rodríguez [2018] recently investigated the issue of consent differences.³³⁴ She was unable to reject the hypothesis that „explicit consent leads to less tracking presence.”

We refrain from investigating a hypothesis and take a different approach. We rank the countries on a scale by their (assumed) strictness of consent implementation by applying our GBMA. First, we zoom in on the *subset of top-20 edges* with a boxplot (Figure 4.6, next page). Second, we will briefly discuss the situation for *all the edges* in Crawl7 (Figure 4.7, next page). Third, we will discuss the issue of differences in implementations of consent mechanisms based on the mutual correlation of the edges per EU country.

I admit that a boxplot for the subset top-20 edges (the three leading companies: Google, Twitter, and Facebook) describes a very small number (20) of distinct edges. Consequently, it provides us only with a (rather) limited view on the situation of implementation differences. However, the main reason for providing such an analysis of the subset top-20 edges, is that it serves as a clear example of how we may interpret the scale.

³³⁴ Turcios Rodríguez [2018, pp. 95–97] by a Generalized Linear Model (GLM). She reported a goodness of fit of the GLM expressed by Akaike Information Criterion (AIC) (8.837171) and McFadden’s Pseudo R² (0.0461).

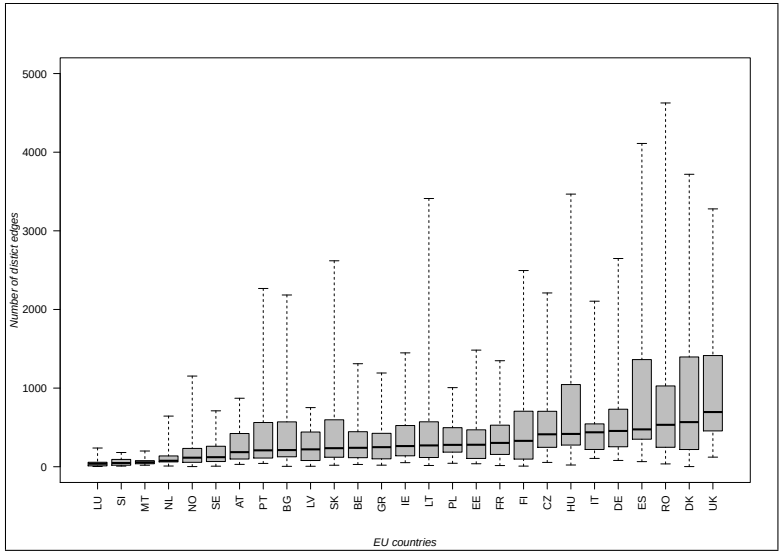


Figure 4.6: Edge boxplot of the subset top-20 edges. Cronbach's alpha: 0.963183 (sample size = 20, number of countries = 27).

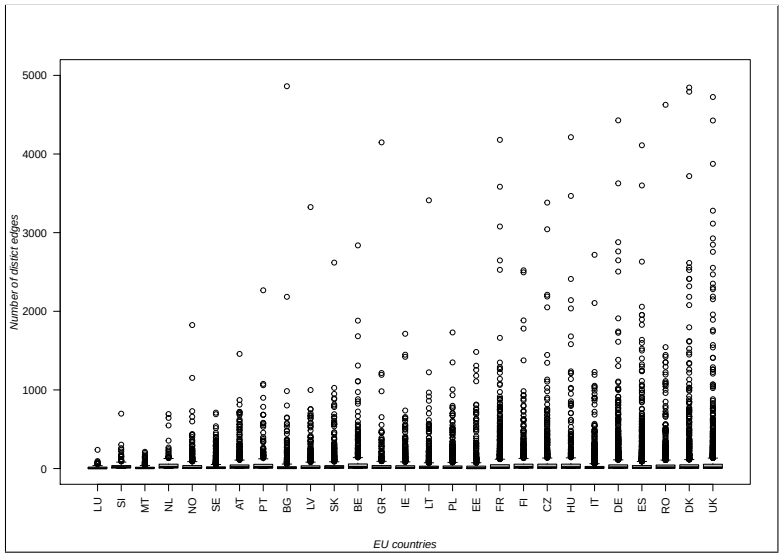


Figure 4.7: Edge boxplot of the full dataset. Cronbach's alpha: 0.8803147 (sample size = 13,789, number of countries = 27).

Below, we discuss (A) a boxplot of the subset of top-20 edges, (B) a boxplot of the full dataset of distinct edges, and (C) the mutual correlation of the edges per EU country.

A: BOXPLOT OF THE SUBSET OF TOP-20 EDGES

In Listing 4.3 we show the corresponding R-code of the boxplot (Figure 4.6).

Listing 4.3: R-code boxplot of the top-20 edges.

```

1 # Selecting the top-20 edges in a dataframe
2 data<-crawl7[2:21,1:27]
3 # Ordering the x-axis of the boxplot of the top-20 edges
4 data_ordered <- data[,c(16,25,18,19,20,24,1,22,3,17,26,2,11,13,
5   15,21,7,10,9,4,12,14,5,8,23,6,27)]
6 boxplot(data_ordered, range=0, srt = 45, las = 2, col =
7   c("grey","grey","grey","grey","grey","grey","grey","grey","
8     grey","grey","grey","grey","grey","grey","grey","grey","
9     grey","grey","grey","grey","grey","grey","grey","grey"),
10   at = c(1,2,3,4,5,6,7,8,9,10,11,12,13,14,15,16,17,18,
11     19,20,21,22,23,24,25,26,27), ylim=c(0,5000))
12 # Adding the x-axis and y-axis labels
13 mtext("EU countries", side = 1, line = 4, cex = 1, font = 3)
14 mtext("Distinct edges", side = 2, line = 4, cex = 1, font = 3)

```

It is noted that the selection of the subset top-20 edges into a dataframe is performed in row 2. The countries are arranged in row 4. Then, the boxplot is formatted (rr. 6–9). Finally, the labels for the x-axis and y-axis are added (rr. 11–12).

On the x-axis in Figure 4.6 we see the countries listed. On the y-axis we see the number of distinct edges. It depicts the five elements of the numerical summary (Tukey [1977]):

- (1) minimum,
- (2) lower-hinge (first quartile),
- (3) median,
- (4) upper-hinge (third quartile),
- (5) maximum.

The sorting order of the countries on the x-axis is by the median value of the number of distinct edges (Listing 4.3, rr. 4–5). The

median value represents a measure for the middle (50%) of the data (top-20 edges). The boxplot does not show statistical outliers.

At first sight, the number of edges (denoted by the 'maximum' line in Figure 4.7 may seem rather high for 'top-20 edges'. Nevertheless, we should remind ourselves that the maximum number of occurrences in a country corresponds with the distinct edges due to the sorting order of the number of countries in which an edge is present (see the clarifying paragraph in Subsection 4.4.4, Observation 1).

B: BOXPLOT OF THE FULL DATASET OF DISTINCT EDGES

Figure 4.7 shows an edge boxplot for *all distinct edges in Crawl7* (a total of 13,789).³³⁵ The outliers are visualized by circles. Furthermore, we can clearly observe that only a few distinct edges have a very high occurrence, where most of the distinct edges occur once or a few times more. We remark that we kept the order of the countries on the x-axis the same as in Figure 4.6 to enable a visual comparison. However, set aside some subtle differences in the order of the countries on the x-axis, we can clearly differentiate between countries with a (relatively) high number of outliers on the right half in comparison with the left half of the boxplot. We leave this task to the reader.

C: MUTUAL CORRELATION OF THE EDGES PER EU COUNTRY

For the mutual correlation of the edges per country, we calculated Cronbach's alpha to test the internal consistency of the scale. However, the R-library Psy [Falissard, 2009] that includes the function for the calculation of Cronbach's alpha does not compute the missing values (N/As) in a dataframe. Therefore, we replaced the missing values in the dataframe by the value zero.³³⁶ The replacement leads to a value of Cronbach's alpha: 0.963183 for the top-20 edges (Figure 4.6) and 0.8803147 for the full dataset in Crawl7 (Figure 4.7). Based on these values for Cronbach's alpha, we may conclude that the internal consistency of the scale is good (i.e., more than adequate).

³³⁵ The corresponding R-code is: 'data<-EU_top_cnt[2:13790,1:27]'.

³³⁶ The corresponding R-code to replace N/A with 0 is: 'data[is.na(data)] <- 0'.

4.5.5 Section conclusions

From the above, we arrive at five section conclusions. They read as follows.

SECTION CONCLUSION 1

A strict implementation of (1) a consent banner or (2) a cookie wall impacts the quantity our research data. If implemented strictly, we simply cannot collect our data. It is remarked that I as a researcher did not resort to using a browser plug-in or script to (explicitly) consent to tracking cookies and similar technologies.

SECTION CONCLUSION 2

However, I collected and retained the contextual research data (Crawl7) at a time when *implied*-consent mechanisms were the technical norm on media websites in Europe. Therefore, we may now conclude as our second subsection conclusion that the year 2016 provided us with a unique window of opportunity to peek into the black box by collecting contextual research data across all EU countries.

SECTION CONCLUSION 3

Our third section conclusion is that the *scale* depicted by the order of the countries on the x-axis in Figure 4.6 represents that European countries are dominated by strict-consent mechanisms on the left end of the scale.

SECTION CONCLUSION 4

Moreover, European countries where implied-consent mechanisms have the upper hand are to be found on the right end of the scale.

SECTION CONCLUSION 5

From the discussion as described in Section 4.5, we may derive our fifth section conclusion, namely that the *lack of strict consent mechanisms* - i.e., cookies are placed before consent has been granted - *in most EU countries allows WPM researchers to analyze*

RTB systems into the depths of their black boxes. By using our Graph-Based Methodological Approach (GBMA) we are now able to identify the relevant network partners. This brings us undoubtedly a step closer to understand how RTB works in practice. Such an understanding will add much value as soon as it comes to a full discussion on the privacy component of RTB.

4.6 GRAPH ANALYSIS OF PARTNER NETWORKS

In this section we present the application of the GBMA to graph analysis of partner networks in RTB systems within the context of European news websites. Through the GBMA we aim to better understand the *network of partners in an RTB system*. It relies on practical work, viz. applying network science algorithms to empirical WPM-data.

In network science, graph algorithms are used to understand and explain (network) phenomena. We refer to Freeman [1977] who summarized the concept of centrality as follows.

„A family of new measures of point and graph centrality based on early intuitions of Bavelas [1948] is introduced. These measures define centrality in terms of the degree to which a point falls on the shortest path between others and therefore has a potential for control of communication. They may be used to index centrality in any large or small network of symmetrical relations, whether connected or unconnected.” [Freeman, 1977, p. 1]

Furthermore, we refer to Schoch [2015] and Takes [2014] as recent studies in network science. In fact, Schoch [2015, p. 12] produced an extensive overview of 108 centrality measures from literature.³³⁷ We consider this overview as a point of reference for our approach to measure the concept of centrality in an RTB system.

Below we discuss the application to the networks of RTB partners of two network science measures, i.e., (1) cluster-edge betweenness for the identification of betweenness clusters (Subsection 4.6.1) and (2) node betweenness for the identification of central nodes (Subsection 4.6.2). The aim of both measures is to calculate the betweenness scores and add these as properties to our

³³⁷ An interactive version of his ‘periodic table of network centrality’ can be found online: URL: <http://schochastics.net/sna/periodic.html> (15 September 2018).

graph (see Subsection 4.4.3). In Subsection 4.6.3 we categorize the nodes in a betweenness cluster into four types of nodes. In Subsection 4.6.4 we provide section conclusions.

4.6.1 Cluster-edge betweenness

Cluster-edge betweenness is a measure that is used for the identification of betweenness clusters. It is a measure based on shortest paths in a graph. It solves an important issue for us: which RTB partners cluster together in an RTB system? The measure helps us to differentiate between RTB systems, i. e., clusters of RTB partners.

Mathematically it can be formulated as follows. Let $g = (V, E)$ be a graph. Let σ_{st} be the total number of shortest paths from s to t and $\sigma_{st}(e)$ be the number of these shortest paths that pass through the edge e . The centrality betweenness for the edge e (henceforth: $C_B(e)$ or cluster-edge betweenness) is

$$C_B(e) = \sum_{s \neq t} \frac{\sigma_{st}(e)}{\sigma_{st}} \quad (4.1)$$

where the sum runs over all s, t pairs with $s \neq t$.

EXAMPLE OF A BETWEENNESS CLUSTER: YOUTUBE-NOCOOKIE

Figure 4.8 shows the results of our attempt to identify subgraphs of RTB partners. We applied cluster-edge betweenness to demonstrate the usefulness of the metric.

Figure 4.8 depicts the cluster-edge betweenness subgraph as a result of visits to two Slovenian news websites, i. e., Dnevnik³³⁸ and Mladina³³⁹. The betweenness cluster consists of 10 nodes (including the central node of the graph) and 17 edges. Both websites contain a YouTube video embedded in a privacy friendly way judging from the FQDN 'www.youtube-nocookie.com'.³⁴⁰

³³⁸ URL: <http://www.dnevnik.si> (30 August 2016).

³³⁹ URL: <http://www.mladina.si> (30 August 2016).

³⁴⁰ For instance, the embedding of the following video: URL: <https://www.youtube-nocookie.com/embed/n7KBIfv4B20> (30 August 2016).

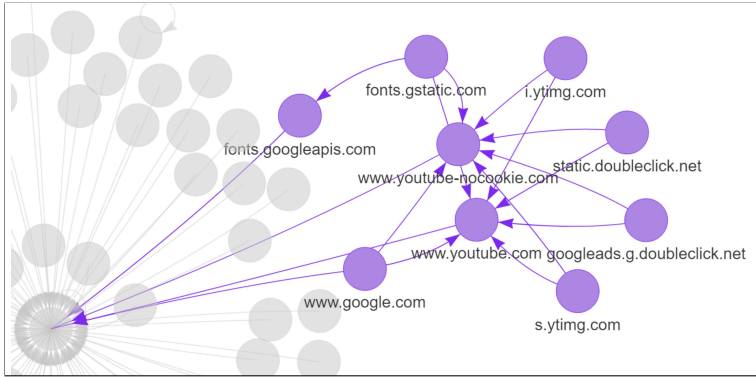


Figure 4.8: Cluster-edge betweenness: Slovenia (SI) with $N=10$ nodes and $E=17$ edges.

The algorithm to compute the measure is based on Newman and Girvan [2004].³⁴¹ We settled for this particular implementation for three practical reasons, i.e., (1) the availability of the source code, (2) its good documentation (which is included in the popular iGraph R-library version 1.0.0 by Csárdi and Nepusz [2006]), and (3) the implementation is similar to the iterative approach proposed by Brandes [2001]. The two computing steps are as follows.

1. Calculate cluster-edge betweenness for every edge.
2. Remove the edge with the highest betweenness value and recalculate cluster-edge betweenness for the remaining edges. Repeat this step.

To calculate the cluster-edge betweenness, we started with an *unweighted* and *undirected* graph while determining the shortest paths algorithmically. There we first remark that taking the *directed* graph as a starting point for the calculation of the shortest paths would lead to (1) a less accurate identification of partner networks and therefore (2) fewer partner networks. With this observation in mind, we compiled an interactive web application with the R-package visNetwork [Thieurmél, 2016]. The web application (Figure 4.8) allowed us (1) to merge the algorithmically identified betweenness clusters as metadata to the nodes of our

³⁴¹ The documentation can be found online: URL: http://igraph.org/r/doc/cluster_edge_betweenness.html (16 September 2018).

directed graph and (2) to interactively explore our property graph depicting partner networks.

PROGRESS OF DISCUSSION: LAUNCH OF THE WHITE HOUSE WEB SITE

I presented, explained, and discussed Figure 4.8 at the Dagstuhl seminar 17162, Online Privacy and Web Transparency [Van Eijk, 2017]. What I knew was that Google offers an option for publishers to turn on a privacy-enhanced mode for embedded YouTube content.³⁴² Good that this has happened. I am indebted to Krishnamurthy [2017], who pointed this out to me that the Google's privacy-enhanced mode exists since 2009, as the direct result of the launch of President Obama's White House Web site.³⁴³

PROGRESS OF DISCUSSION: ENFORCEMENT BY THE SLOVENIAN DATA PROTECTION AUTHORITY

Furthermore, we remark that we measured HTTP requests to 'www.youtube-nocookie.com' only in Slovenia (August 2016). Here I learned from a freedom of information request to the Slovenian Data Protection Authority (DPA) [Burnik, 2017] that the use of the privacy option by the two publishers was the direct result of enforcement actions by the DPA.³⁴⁴

In fact, the Slovenian DPA had enforced in 12 cases where publishers embedded YouTube in such a way that a visit to their website resulted in tracking cookies being stored in the end-user's browser without their consent. Out of 12 cases, three were related to news websites, i.e., (a) Dnevnik, (b) Mladina, and (c) Demokracija and five cases were related to public sector websites (municipalities). At this point, we may conclude that strong enforcement by the DPA explains why Slovenia (SI) is positioned on the left end of our scale (Figure 4.7). In fact, strong enforcement is the main reason for the implementation of strict-consent mechanisms (Definition 4.14) in Slovenia.

³⁴² URL: <https://support.google.com/youtube/answer/171780?hl=en> (18 march 2018).

³⁴³ See, e.g., URL: <https://www.cnet.com/news/white-house-acts-to-limit-youtube-cookie-tracking/> (18 March 2018).

³⁴⁴ The Slovenian DPA is the competent authority for enforcement of Article 5(3) EPD.

4.6.2 Node betweenness

In this subsection we aim to identify nodes that are central in a betweenness clusters of RTB partners. The concept of node betweenness was, as cited in Freeman [1977], introduced by Bavelas [1948]. The concept was named ‘point centrality’ and based on the intuitive notion that „when a particular person in a group is strategically located on the shortest communication path connecting pairs of others, that person is in a central position”.³⁴⁵ See also, similar conceptualizations of node betweenness by Cohn and Marriott [1958], Shaw [1954], and Shimbel [1953].³⁴⁶

Node betweenness solves an important issue for us: which are the dominant companies within a network of RTB partners? Node betweenness helps us to differentiate between the companies.

Similar to the formulation of cluster-edge betweenness (Equation 4.1), node betweenness can be formulated mathematically. Let $g = (V, E)$ be a graph. Let σ_{st} be the total number of shortest paths from s to t and $\sigma_{st}(v)$ be the number of these shortest paths that pass through the vertex v . The centrality betweenness for the vertex v (henceforth: $C_B(v)$ or node betweenness) is:

$$C_B(v) = \sum_{s \neq v \neq t \in V} \frac{\sigma_{st}(v)}{\sigma_{st}} \quad (4.2)$$

where the sum runs over all s, t pairs with $s \neq t \neq v \in V$.

In Figure 4.9 we applied node betweenness as a weight to the nodes in a subgraph.³⁴⁷ This way, the size of a node becomes an indication of the notion of importance of the node. The larger the node, the more important its role is in relation to the other nodes (see Figure 4.9, i. e., (V_1^*) ‘www.youtube-nocookie.com’ and (V_2^*) ‘www.youtube.com’ in comparison with the remaining nodes in the cluster).

Finally, we provide two remarks. First, we remark that the betweenness score for the center of the graph (‘crawled.io’) has been curated (set to zero) because our focus is on the nodes in the

³⁴⁵ As cited in Freeman [1977].

³⁴⁶ Ibid.

³⁴⁷ We applied a node-betweenness algorithm from the popular iGraph R-library version 1.0.0 by Csárdi and Nepusz [2006] to our data. The implementation of the algorithm is based on Brandes [2001]. The documentation can be found online: URL: <http://igraph.org/r/doc/betweenness.html> (16 September 2018).

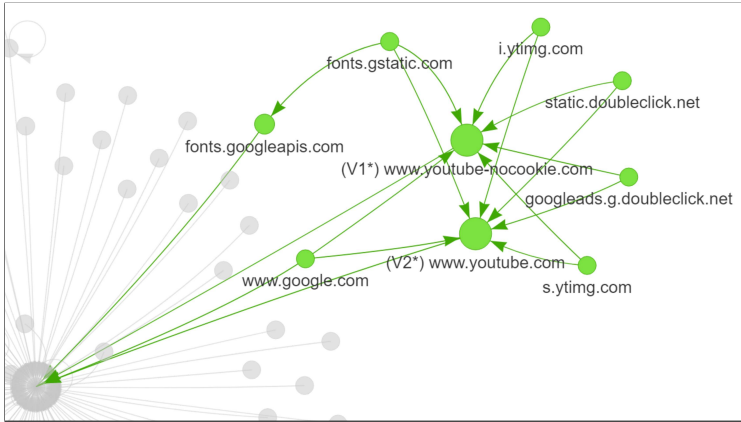


Figure 4.9: Central nodes: Slovenia (SI). Weighted nodes for the betweenness cluster YouTube with $N=10$ nodes and $E=17$ edges, $g = (V, E)$, and two central nodes V_1^* : 'www.youtube-nocookie.com', V_2^* : 'www.youtube.com'.

clusters. Second, we remark that the reason why the center node in Figure 4.9 seems a large grey node is due to the points of the arrows surrounding it.

4.6.3 Categorization into four types of nodes

In this Subsection we will categorize the occurrences of nodes into four types, viz.

- (1) central node,
- (2) satellite node,
- (3) bridging node, and
- (4) interconnecting node.

Below we discuss all four of them, with emphasis on the bridging and interconnecting types.

Barrat, Barthélemy, and Vespignani [2007, p. 3] remarked that *central* nodes are part of a higher number of shortest paths within the network than *satellite* (or peripheral) nodes.³⁴⁸ Moreover, they emphasized the crucial role for central nodes and for *bridging*

³⁴⁸ The full quotation is as follows: „Central nodes are therefore part of more shortest paths within the network than peripheral nodes.”

nodes to connect different regions of the network by acting as bridges.³⁴⁹

In Figure 4.10 we provide an example of the role for bridging nodes. It depicts an annotated cluster with three bridging nodes:

- (v1*) ('fonts.gstatic.com'),
- (v2*) ('fonts.googleapis.com'), and
- (v3*) ('www.google.com').³⁵⁰

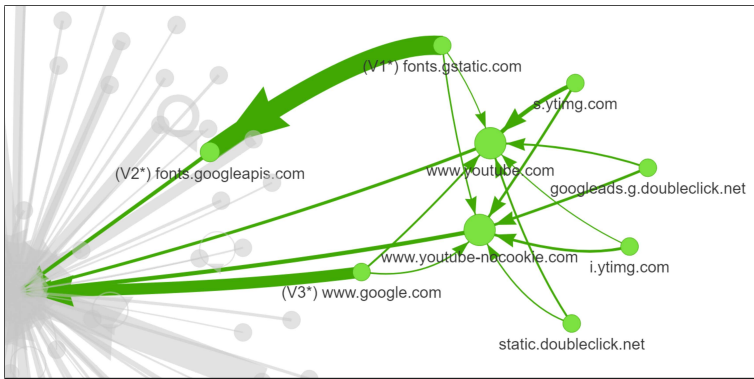


Figure 4.10: Annotated cluster: (Slovenia, SI). Weighted nodes and weighted edges for the betweenness cluster YouTube with $N=10$ nodes and $E=17$ edges, $g = (V, E)$, $E_1: V_1^* \times V_2^*$ ('fonts.gstatic.com' -> 'fonts.googleapis.com'), and $E_2: V_3^* \times V_{io}^*$ ('www.google.com' -> 'crawled.io').

The edge annotation tells us something about the referrer flow of data, which originated from visiting the webpages depicted in the refactored node V_{io}^* 'crawled.io'. That is the bridging role of the nodes. In fact, the HTTP headers for this subgraph contained more referrals from $V_1^* \times V_2^*$ (E_1) and $V_3^* \times V_{io}^*$ (E_2) than the remaining edges. We remark that the bridging function contributes to our understanding of the role of the central nodes 'www.youtube.com' and 'www.youtube-nocookie.com' in the cluster.

349 The full quotation is as follows: „by considering solely the [betweenness] degree of a node we overlook that nodes with small [node-betweenness] degree may be crucial for connecting different regions of the network by acting as bridges.”
350 We added the number of edges as a weight to the edges to highlight the node’s role, i.e., $E_1: V_1^* \times V_2^*$ and $E_2: V_3^* \times V_{io}^*$ ('crawled.io').

Based on the clarification provided above, we propose a fourth and special type of bridging node, i.e., the *interconnecting* node. Where a bridging node connects different regions *within* our cluster, an interconnecting node acts with the crucial role of connecting between *different* clusters. Therefore, we propose to categorize the (remaining) nodes of a graph $g = (V, E)$ into the four node types mentioned at the beginning.

Below we provide intuitive definitions for each of them, i.e., central node (henceforth: V_c , Definition 4.16), satellite node (henceforth: V_s , Definition 4.17), bridging node (henceforth: V_b , Definition 4.18), and interconnecting node (henceforth: V_{ic} , Definition 4.19). Moreover, we use V_{io} , meaning a representation of the refactored node of a graph g .

DEFINITION 4.16: CENTRAL NODE. Let us consider a refactored referrer graph denoted by $g = (V, E)$ with a betweenness-centrality cluster $G(\alpha) = (V_\alpha, E_\alpha)$ and V_{io} representing the refactored node of graph g , i.e., the webpages visited in a crawl. A *central* node is defined to be a node V_{c_i} with a high node-betweenness score in comparison to the remaining nodes in the same cluster $G(\alpha)$; with $V_{c_i} \in V_\alpha$ and $V_{c_i} \neq V_{io}$.

DEFINITION 4.17: SATELLITE NODE. Let us consider a refactored referrer graph denoted by $g = (V, E)$ with V_{io} representing the refactored node of graph g , $G(\alpha) = (V_\alpha, E_\alpha)$ representing a betweenness cluster α , $V_c(\alpha)$ representing one or more central nodes, and $V_b(\alpha)$ represents zero or more bridging nodes. A *satellite* node is defined to be a node V_{s_i} connected to $V_c(\alpha)$ but not to V_{io} by its shortest path ($V_{s_i} \in V_\alpha$ with $V_{s_i} \neq V_{io}$, $V_{s_i} \neq V_c(\alpha)$, $V_{s_i} \neq V_b(\alpha)$).

DEFINITION 4.18: BRIDGING NODE. Let us consider a refactored referrer graph denoted by $g = (V, E)$ with V_{io} representing the refactored node of graph g , $G(\alpha) = (V_\alpha, E_\alpha)$ representing a betweenness cluster α , $V_c(\alpha)$ representing one or more central nodes, and $V_s(\alpha)$ represents zero or more satellite node. A *bridging* node is defined to be a node V_{b_i} connected to V_{io} and $V_c(i)$ by its shortest path ($V_{b_i} \in V_\alpha$ with $V_{b_i} \neq V_{io}$, $V_{b_i} \neq V_c(\alpha)$, $V_{b_i} \neq V_s(\alpha)$).

DEFINITION 4.19: INTERCONNECTING NODE. Let us consider a refactored referrer graph denoted by $g = (V, E)$ with multiple betweenness-centrality clusters $\Gamma(i) \in G, G = ((V_\alpha, E_\alpha), (V_\beta, E_\beta), \dots, (V_N, E_N))$, with $i \in N$ and V_{io} representing the refactored node of graph g . $V(\alpha)$ represents a node in a subgraph $\Gamma(\alpha)$ with $V(\alpha) \neq Vc(\alpha)$, and $Vc(\beta)$ represents a central node in subgraph $\Gamma(\beta)$. An *interconnecting* node is defined to be a node $V(\alpha)$ connected to $Vc(\beta)$ via edge $E(\alpha)$: $V(\alpha) \times Vc(\beta)$.

EXAMPLE OF THE FOUR NODE TYPES: LUXEMBOURG (LU)

To illustrate our approach, we provide an example of all four node types in Table 4.4 and the corresponding Figure 4.11. The example depicts our categorization of node types from the data in Crqw17 from Luxembourg (LU), which is situated on the left hand of the scale (Figure 4.6).

Table 4.4: Luxembourg (LU): central nodes (9), satellite nodes (44), bridging nodes (7), and interconnecting nodes (2), and $E=81$ edges.

CLUSTER	CENTRAL NODES	SATELLITE NODES	BRIDGING NODES	INTERCONNECTING NODES
Brightcove	V_{c_1}	$V_{s_1}-V_{s_{11}}$	$V_{b_1}-V_{b_3}$	-
Twitter	V_{c_2}	$V_{s_{12}}-V_{s_{14}}$	-	-
Youtube	V_{c_3}	$V_{s_{15}}-V_{s_{18}}$	-	V_{ic_1}
Spotify	V_{c_4}	$V_{s_{19}}-V_{s_{21}}$	V_{b_4}	-
Google	V_{c_5}	$V_{s_{22}}$	V_{b_5}, V_{b_6}	V_{ic_2}
DoubleClick	V_{c_6}	$V_{s_{23}}$	-	-
Facebook	V_{c_7}	$V_{s_{24}}$	-	-
Cloudflare	V_{c_8}	-	V_{b_7}	-
'crawled.io'	V_{io}	$V_{s_{25}}-V_{s_{44}}$	-	-

We remark that the name for each cluster is denoted in Table 4.4 in the first column 'cluster' by the organization representing the central node in a betweenness cluster. Then we provide the node types in the second column 'central node', the third column 'satellite node', the fourth column 'bridging node', and the fifth

column 'interconnecting node'. The values in the table, e.g., V_{c_1} , V_{s_1} , V_{b_1} , or V_{ic_1} , correspond to the labels of the nodes in the graph (Figure 4.11).

Furthermore, we note that we did not remove the loops that connect a node to itself in Figure 4.11. The self-referencing loops represent valuable metadata, i.e., a HTTP referrer header or a HTTP location header pointing to the same origin (Subsection 4.4.3).

Obviously, it would take too long to discuss each country of our scale one by one. For more insight into the intricacies we provide in Appendix D a few, somewhat more complex examples of partner networks in Denmark (DK), which is situated on the right hand of the scale.³⁵¹ The aim of the examples is to zoom in on five leading partner networks and their interconnections:

- (1) the Rubicon Project (Figure D.2, p. 269),
- (2) Adform (Figure D.3, p. 270),
- (3) Adspine (Figure D.4, p. 271),
- (4) AppNexus (Figure D.5, p. 272), and
- (5) Google (Figure D.6, p. 273).

PROGRESS OF DISCUSSION: EIGENVECTOR CENTRALITY

Clearly, network science has produced many other useful algorithms. In fact, some algorithms may also be useful within the context of RTB. If we were to identify a category of algorithms for further research, we recommend investigating, e.g., eigenvector centrality [Bonacich, 1987].³⁵² We note that eigenvector centrality is an alternative to the calculation of a node betweenness value. Eigenvector centrality values are valuable metadata in our context since we can add the measure as a weight to the nodes in our refactored referrer graph. Eigenvector centrality is calculated by assessing the connectedness of a node to parts of the network with the greatest connectivity. Nodes with a high eigenvector are RTB-network partners with many connections to other partners (with many connections themselves). Such an interconnection is a property of, e.g., an ad exchange (RTB-step 4 - RTB-step 6).³⁵³

³⁵¹ Denmark (DK) with $N=719$ nodes, $E=1,348$ edges, and $C_B=33$ clusters (see Figure D.1, p. 267).

³⁵² See n. 337.

³⁵³ See Subsection 4.3.9.

4.6.4 *Section conclusions*

Based on the above analyses, we may draw one main section conclusion and four specific section conclusions.

MAIN SECTION CONCLUSION

Cluster-edge betweenness and node betweenness solve at least four important issues for us.

SECTION CONCLUSION 1

Cluster-edge betweenness provides a direct answer to the question: which companies cluster together in a partner network? Cluster-edge betweenness helps us to identify partner networks empirically. So, we may conclude that the nodes in a betweenness cluster represent partners in a network.

SECTION CONCLUSION 2

Node betweenness helps us to identify the main actors within a partner network. Here, we may conclude that nodes with a high node-betweenness value represent partners with a central role in the RTB network. It is crucial to remark that the unweighted and undirected graph is the basis for both betweenness measures.

SECTION CONCLUSION 3

We may conclude that categorization of the nodes into four types (i. e., central node, satellite node, bridging node, and interconnecting node) helped us to answer the two important questions: (1) how do the partners relate within a partner network? and (2) how do partners relate in between partner networks? By categorizing all actors of a partner network into node types, we can now build a fully understandable picture of the cascading HTTP requests and HTTP responses triggered by the ads. Precisely in this point, adding metadata as annotations to the referrer graph comes in (i. e., (1) node betweenness as a weight to the nodes, (2) directed edges as source of information and (3) adding the number of occurrences as a weight to the edges). The annotated cluster

helps us to understand the cascading flow of raw HTTP data in a very comfortable and informative way.

SECTION CONCLUSION 4

We may conclude that the interconnectedness in our graph *inside* a partner network and *between* partner networks helps us to differentiate quite precisely between clusters in the way that they are either ad related and that others are non-ad related (see Definition 4.1). The difference is important, because (obviously) not all HTTP activities in the graph can be attributed to the ads.

4.7 CHAPTER CONCLUSIONS

In this chapter we presented five findings of our Graph-Based Methodological Approach (GBMA) to the categorization of RTB systems. They read as follows.

- (1) Digital media in Europe is to be seen as a contextual data source (Subsection 4.4.1).
- (2) Graph refactoring can be performed by grouping the origin of all digital news items visited on a single node Definition 4.13).
- (3) The identification of partner networks can be performed by applying a cluster-edge betweenness algorithm to the graph (Section 4.6.1).
- (4) The identification of the main actor(s) in a partner network can be performed by applying a node betweenness algorithm to the graph (Section 4.6.2).
- (5) The categorization of graph nodes can take place into four types (a) central node, (b) satellite node, (c) bridging node, and (d) interconnecting node (Subsection 4.6.3).

All in all, we identified differences in consent implementations for tracking cookies in European countries. To do so, we ranked the countries on a scale (Section 4.5). From the results, we arrive at five chapter conclusions. The first three are conclusions following from the analysis of the prevailing differences in consent in Europe. Chapter Conclusion 4 and 5 follow from the graph analysis.

CHAPTER CONCLUSION 1

The leading RTB companies are omnipresent in most EU countries. This prevalence gives them a strategic *cross-border* position. The capabilities in terms of data collection of the leading RTB are amplified if we take (cross-border) partner networks into account.

CHAPTER CONCLUSION 2

European countries are dominated by strict-consent mechanisms on the left end of the scale.

CHAPTER CONCLUSION 3

European countries where implied-consent mechanisms have the upper hand are to be found on the right end of the scale.

The Chapter Conclusions 2 and 3 are supported by applying the GBMA with the network science algorithms applied to our contextual research data.

CHAPTER CONCLUSION 4

We showed that (1) cluster-edge betweenness (Section 4.6.1) and (2) node betweenness (Section 4.6.2) help us in understanding the interconnectedness of the ad-technology companies. We may therefore conclude that the *combined* application of both network science algorithms to empirical WPM-data enables us to categorize closely interconnected dataflows.

CHAPTER CONCLUSION 5

The interconnection in a betweenness cluster is caused by the data flows of the companies themselves due to their specializations in ad technology (Subsection 4.1.5).

The fifth conclusion is our main chapter conclusion. The reason is that it shows us the prevailing reason why the application of network science algorithms is effective. As a result of the specializations, the dataflow is functional and contains structural metadata which we use to build and annotate our graph (Subsection 4.6.3).

By modeling the context of European news websites with specific small-data elements we can now address RQ2 (see Section 4.8). Our GBMA to provide an answer is based on the application of network science to WPM (see Section 4.6). It is a new approach in comparison with other web-tracking visualization tools (see Section 2.7).

A SERIOUS LIMITATION

Finally, we remark that a serious limitation of our approach is, that *if* the graph becomes too large, the categorization of partner networks becomes less precise. In very large graphs the algorithm starts to combine partner networks. For instance, the combined refactored graph of all 27 countries with $N=5,988$ nodes and $E=12,703$ edges (henceforth: EU graph), will result in the identification by cluster-edge betweenness which is less precise. An example may elucidate this limitation. We identified betweenness clusters in the separate graphs of each country. Germany is leading with 40 identified clusters, followed by Spain (29), Czech Republic (27), and United Kingdom (27).³⁵⁴ The total is 512 clusters in 27 countries.³⁵⁵ However, the combined EU graph contains 160 clusters.³⁵⁶ How can we handle this difference?

In Observation 2 (Subsection 4.4.4) we remarked that the Rubicon Project puts its partner network in a good position to track end-users across news websites in 25 EU countries. Closer inspection of the EU graph shows that the degree of overlap between the partner networks depends on the scale at which researchers intend to analyze the data. Obviously, there is overlap (through the lens of the network science algorithm) of partner networks across countries. To put the limitation of our approach into perspective, we remark that marketers tend to operate from a somewhat inextricable link with the end-users in a country. In other words, the marketing message is in general tailored to end-users within a country. Therefore, we remark that our approach is well suited

354 The number of identified clusters per country is as follows: DE (40), ES (29), CZ (27), UK (27), IT (26), DK (25), PL (25), BE (24), HU (23), LT (21), FR (19), RO (19), FI (18), LV (18), PT (18), EE (17), SE (17), BG (15), NO (15), SK (15), IE (14), MT (13), NL (12), GR (11), AT (10), LU (8), SI (6).

355 We corrected for the 'crawled.io'-cluster and 251 unconnected satellite nodes.

356 We identified 160 in the EU graph (after correction for 'crawled.io' and 82 satellite nodes).

for the purpose of categorization of partner networks on a country level, e. g., within the context of European news websites.

4.8 AN ANSWER TO RESEARCH QUESTION 2

As stated above, we are now able to answer RQ2: what are the emerging characteristics of the graph that is fit for graph analysis? To answer this question, we will present four views on RTB systems: an empirical view, a theoretical view, a legal view, and a societal view.

EMPIRICAL VIEW: NETWORK OF PARTNERS

We started off with an *empirical* view on the network of partners, by zooming in on a consolidation over a period of three years caused by acquisitions, mergers, and strategic partnerships (Section 4.1). What partners in an RTB system do is strengthening their positions. We demonstrated growth in partner networks by a longitudinal analysis of the Rubicon Project partner network over a period of almost six years. We concluded that each network of partners (1) relies on its own targeted-advertising framework and (2) has its unique way of tracking end-users (for various purposes).

We defined an RTB systems to be a network of partners enabling big data applications within the organizational field of marketing to improve sales by real-time data-driven marketing and personalized (behavioral) advertising (Section 4.2).

THEORETICAL VIEW: NETWORK OF EIGHT KEY BUILDING BLOCKS

Next, we presented a *theoretical* view of RTB systems (Section 4.3). It consists of eight key building blocks and a brief description of the nine RTB-steps. We concluded that partnerships enable the real-time integration of RTB systems. Moreover, the interconnection of RTB systems and their partners with standardized RTB protocols creates the infrastructure for RTB.

LEGAL VIEW: REQUIREMENTS FOR CONSENT

We developed a *legal* view by linking our theoretical view to an empirical view on partner networks (Section 4.4). We based our

view on the analysis of the data from news websites from 27 EU countries in our dataset. We showed a difference in implementations of consent mechanisms implementing the legal requirement for consent across European countries. The reason for the differences can be traced back to (1) the legal norm being enshrined in a EU directive, which left EU member states (some) room transposing in national law the legal requirements for consent for tracking cookies; and (2) a difference in effective enforcement by the competent authorities for article 5(3) of the EPD. We concluded that strong enforcement is the main reason for the implementation of strict-consent mechanisms in Slovenia. We described the landscape in August 2016 and provided a methodology to gauge the situation of differences in consent implementations.

SOCIETAL VIEW: THE CONTEXT OF DIGITAL MEDIA IN EUROPE

Finally, we developed a *societal* view on graph analysis of partner networks in RTB systems within the context of digital media in Europe (Section 4.6). It is based on the application of algorithms from the domain of network science to empirical WPM-data. We concluded that the nodes in a mathematically determined betweenness cluster represent partners in a network. Furthermore, we concluded that nodes with a high node-betweenness value represent partners with a central role in the RTB network. These observations led to the view that categorization of the nodes into four types, i. e., (1) central node, (2) satellite node, (3) bridging node, and (4) interconnecting node; helped to better understand the cascading flow of raw HTTP data. All in all, we came to the conclusion that the graphical representation of the interconnect-edness (1) inside a partner network and (2) between partner networks helped us to precisely differentiate between clusters that are ad related and other non-ad related.

THE PROPERTIES OF THE GBMA

Here we reiterate that our graph-based approach consists of seven properties:

- (1) contextual data,³⁵⁷
- (2) modelling a directed graph with referrer metadata,

³⁵⁷ In contrast to crawling a list of the top (1m) sites (see Section 3.1).

- (3) redirection metadata, and
- (4) graph refactoring; and
- (5) the application of a cluster-edge betweenness algorithm and
- (6) a node betweenness algorithm to the graph, and
- (7) the categorization of graph nodes into four node types.

The properties enable us to analyze the emerging characteristics of the graph. The main improvement of capabilities for categorization of RTB systems is the fact that the *combined* application of these properties enables us to categorize closely interconnected HTTP dataflows.

THE ANSWER IS IN THE CONSEQUENCES

Our GBMA showed that the difference in specializations by companies is the main reason why network science algorithms work well on refactored subgraphs of websites containing RTB ads. The interconnection in a betweenness cluster is caused by the data flows of the companies themselves due to their specializations in ad technology. As a result of the specializations, the dataflow is functional and contains structural metadata which we use to build and annotate our graph. This is the answer to RQ2.

THE FUTURE

The new and deep insight into partner networks and their interconnections is important against the background of the changes in the *legal* framework for online marketing in Europe. With the delay of the legislative process of the EPR (the *lex specialis* to the GDPR) and the GDPR in force, the current adequate insight into the intricacies triggers the intriguing question: how fast will the landscape of consent mechanisms shift toward strict implementations?

5

Conclusions

In this chapter, we provide a graph-based answer to our research questions RQ1, RQ2, and to the problem statement. In Section 5.1 we summarize the two research questions and the answers provided in Chapter 3 and Chapter 4. In Section 5.2 we formulate our answer to the PS and draw a general conclusion from our research work. In Subsection 5.3 we discuss Step 8 of our research methodology, viz. evaluation of the framework. Finally, in Section 5.4 we conclude the chapter by indicating directions for further research.

5.1 ANSWERS TO THE RESEARCH QUESTIONS

RESEARCH QUESTION 1: How do we move from data collection to graph analysis?

The answer to our first research question is derived from Chapter 3. Addressing RQ1 we presented a graph-based approach with small-data models based on metadata from the HTTP header. The properties of the nodes and edges in the small-data model correspond with the properties of the metadata as defined in our research environment consisting of data stored on websites.

The main improvement of the research methods and the capabilities for analysis is the fact that a small-data model for web tracking allows us to apply a variety of network science algorithms to the data. The Graph-Based Methodological Approach (GBMA) allows us to deepen our understanding of what actually happens when we visit websites containing RTB advertisements.

RESEARCH QUESTION 2: What are the emerging characteristics of the graph that is fit for graph analysis?

The answer to our second research question is derived from Chapter 4. Addressing this question, we presented four views on RTB systems. We started with an *empirical* view on the network of partners (View 1) and we may conclude that (1) each network of partners relies on its own targeted-advertising framework and (2) has its unique way of tracking end-users for various purposes.

Next, we presented a *theoretical* view of RTB systems (View 2) consisting of eight key building blocks. We may conclude that partnerships enable the real-time integration of RTB systems. Moreover, the interconnection of RTB systems and their partners with standardized RTB protocols creates the infrastructure for RTB.

We developed a *legal* view by linking our theoretical view to an empirical view on partner networks (View 3). We described the landscape in August 2016. We based our legal view on the analysis of a contextual dataset comprised of news items from news websites from 27 European countries. Furthermore, we provided the Graph-Based Methodological Approach (GBMA) to gauge the situation of differences in implementations of consent mechanisms for (tracking) cookies.

Finally, we showed our *societal* view on graph analysis of partner networks in RTB systems (View 4). The graphical representation of the interconnectedness (1) inside a partner network and (2) between partner networks helped us to precisely differentiate between clusters that are ad related and other non-ad related.

In summary, we may conclude that RTB as a technology relies on dynamic networks of partners to enable an ad. The interconnection in a betweenness cluster is caused by the data flows of the companies themselves due to their specializations in ad technology. As a result of the specializations, the dataflow is functional and contains structural metadata which we use to build and annotate our picture.

5.2 ANSWER TO THE PROBLEM STATEMENT

Below we provide an answer to the problem statement based on the flow of arguments throughout the thesis and the eight steps of our research methodology. The primary focus of our investigation was: to address on the gap between (1) what end-users think is

going on and (2) what is actually happening? The two issues led us to formulate a problem statement connecting privacy and RTB.

PROBLEM STATEMENT: *To what extent can we measure the privacy component connected to Real-Time Bidding?*

The answer to the problem statement is as follows. The network of RTB partners measured from the perspective of an end-user is the start of an extensive (say disruptive) RTB dataflow. Based on our research findings we arrive at four conclusions. They are different and address the following topics: (1) specializations, (2) visualization and annotation, (3) limitations, and (4) privacy impact.

CONCLUSION 1: SPECIALIZATIONS

We may conclude that the distinction between RTB systems is based on the specializations in ad technology of the partner companies. As a result of the specializations, the dataflow is functional and contains structural metadata.

CONCLUSION 2: VISUALIZATION AND ANNOTATION

We may conclude that we can visualize and annotate the interconnection in a partner network with a betweenness cluster algorithm.

CONCLUSION 3: LIMITATIONS

We may conclude that we cannot measure the full dataflow since most of the RTB functions are performed through sharing data outside the sphere of the browser.

CONCLUSION 4: PRIVACY IMPACT

We may conclude that due to a lack of metadata in current RTB implementations, e. g., about the identity of a partner and the purpose of a UID, we are rather limited in differentiating between UIDs. There are UIDs with a high impact on the privacy of an end-user (relating, e. g., targeting or retargeting cookies) and those with a smaller impact (e. g., UID cookies for campaign measurement).

5.3 FUNDAMENTAL DATA PROTECTION RIGHTS

As briefly addressed in Section 1.2, the privacy of an end-user should be regarded in conjunction with fundamental rights. In Recitals 1–13 GDPR, the European legislator explains the relation between the principles, the legal norms in the GDPR, and the fundamental right to the protection of personal data.

Furthermore, the European legislator explains the boundaries to the concept of protecting personal data in Recital 4 GDPR:

„The right to the protection of personal data is not an absolute right; it must be considered in relation to its function in society and be balanced against other fundamental rights, in accordance with the principle of proportionality.“

MULTIPLE (JOINT) CONTROLLERS

Our answer to the problem statement leads us - with the fundamental rights in mind - to two intriguing questions for future research: (1) is RTB as a technology disproportionate to its intended use? and (2) should RTB be prohibited if (a) data leakage to unknown (joint) controllers cannot be prevented and (b) leaked data is subsequently processed for unknown purposes? Whatever the answer, we take the stance that, in principle, end-users should be able to exercise their fundamental data-protection rights.³⁵⁸

Art. 29 WP provided recently a guidance on the information requirement in a situation of multiple (joint) controllers (as defined in Article 4(7) GDPR).³⁵⁹

(...) „Art. 29 WP notes that in a case where the consent sought is to be relied upon by multiple (joint) *controllers* or if the data is to be transferred to or processed by other controllers who wish to rely on the original

³⁵⁸ We remark Leenes and Kosta [2015], who also questioned whether the renewed informed consent framework of the GDPR can change the current status quo in web tracking.

³⁵⁹ Article 4(7) GDPR reads as follows: „controller means the natural or legal person, public authority, agency or other body which, *alone or jointly with others, determines the purposes and means of the processing of personal data*; where the purposes and means of such processing are determined by Union or Member State law, the controller or the specific criteria for its nomination may be provided for by Union or Member State law.“ (emphasis added)

consent, these organizations should all be named. *Processors* do not need to be named as part of the consent requirements, although to comply with Articles 13 and 14 of the GDPR, controllers will need to provide a full list of recipients or categories of recipients including processors. To conclude, Art. 29 WP notes that depending on the circumstances and context of a case, more information may be needed to allow the *data subject* to genuinely understand the processing operations at hand." [Article 29 Working Party, 2018, WP259 rev.01] (emphasis added)

Below we provide an example on remnant inventory on a publisher's website. We aim to illustrate the (joint) responsibility with respect to the GDPR of an *advertiser* when resorting to RTB for prospects (potential new customers). The new customers have expressed an interest by visiting a product page on the corresponding website. We remark that the use of RTB technology for this purpose is a (relatively) new phenomenon in some private sectors that are known for processing data of a sensitive nature (e.g., insurance companies and financial service providers).

EXAMPLE: REMNANT INVENTORY

We already know that publishers largely depend on RTB advertising based on remnant inventory. Visiting a product page of a *non-premium advertiser* (Context A) may lead to an ad on a publisher's website (Context B). The publisher shows ads to its readers of its *premium advertisers* wishing to market them to the newspaper's audience. However, a visit to the publisher's website may also lead to an ad shown to the end-user based on remnant inventory. In other words, a (re)targeted ad from a *non-premium advertiser* based on data about the user collected in Context A may be shown to the end-user in the publisher's Context B. We may conclude that due to a dependency on remnant inventory in the RTB landscape, RTB ads can follow prospects when they navigate from one website to another website.

COLLECTION OF RTB DATA

The interesting point we would like to convey with the above example is, that the non-premium advertiser's product page - where

the RTB data originates from - does not have to include RTB *ad space* itself when resorting to RTB technology. In absence of RTB-ad space, the advertiser may restrict the use of RTB technology to partners that specialize in *collecting* data about the end-user for the purpose of marketing them to other websites on the web. Such a restriction is fully understandable since the advertiser does not want to run the risk of ads from competitors on its own website.

LACK OF USER CONTROL

The example above leads to our view that an advertiser is (most likely) a (joint) data controller with respect to the GDPR for the (re)targeted advertisement (in Context A).³⁶⁰ As a result the burden is - among other (joint) data controllers - on the publisher as well as on the advertiser to demonstrate the legal basis for the ad. Moreover, the advertiser is accountable for the collection of the end-user's data for the purpose of the (re)targeted RTB campaign in Context A and the (subsequent) processing of his data by tens, possibly hundreds of RTB parties. Moreover, the end-user should be able to exercise his fundamental rights, e. g., revoking consent to prevent further data collection and use.

We remark that current RTB frameworks tend to operate in the interest of the *ad exchange* and (in most cases) provide for information about the origin of the (behavioral) ad to the end-user when the ad is rendered. Conversely, we may conclude that current RTB frameworks do not (tend to) operate in the interest of the *advertiser* and the *end-user*, i. e., by providing a mechanism to provide for revocation of consent (in Context A).

ENFORCEABILITY OF DO NOT TRACK MECHANISMS

The legislative intervention by the European legislator with the GDPR is promising, and simultaneously complex. At the time of writing it is unclear whether consent expressed by using the appropriate technical settings of a software will be enforceable by DPAs. Since the GDPR-model of transparency and consent raises technical and legal questions when applied to the dynamic nature of RTB networks, we advocate to include a provision in the EPR on the enforceability by DPAs of Do-Not-Track expressions through

³⁶⁰ To the extent that the collected RTB data qualifies as personal data as defined in Article 4(1) GDPR.

modern browsers. The EPR will further clarify the GDPR by translating its principles in specific rules. Such a provision would not only strengthen the transparency, user control, and trust (see Recital 66 EPD, Article 21(5) GDPR, and the article on consent in the (draft) EPR), but the provision is - we advocate - also needed to catalyze the change to bring back the principles of lawfulness, fairness, and transparency to RTB (Article 5(1) a-f and 5(2) GDPR).

5.4 FURTHER RESEARCH

Further research in the field of WPM may focus on legal and technical aspects of RTB. Below, we propose three new directions for further research.

1. PROTOTYPE A REPORTING PLUG-IN

Now that we have virtually developed a new technology (the GBMA) to differentiate between RTB networks, a next step could be to develop a prototype of a plug-in for GDPR compliance reporting. Prototyping could help publishers to better understand the RTB technology which they embed in their webpages. In this thesis we have learned that data leakage to unknown parties who may process the RTB data for unknown purposes has become a risk for the publisher. Especially because accountability is a key element in the GDPR. This means that the burden of proof with regards to the lawfulness of processing of online identifiers for RTB purposes rests (first and foremost) upon the publisher.

2. DEVELOPMENT OF NEW A MEASURES

Further research into partner networks should include the development of new a measures. For instance, new a measure may be derived from the node types, e. g., the interconnecting node (Definition 4.19). In fact, the proportion of edges connected to a partner network of an interconnecting node can be compared with the size of a partner network (cf. Blondel, Guillaume, Lambiotte, and Lefebvre [2008, Figure 4, p. 7]). Such a measure would allow us to zoom in (1) on interconnecting nodes in large(r) referrer graphs and (2) on the purpose(s) of the processing of the data flowing through the nodes.

3. INVESTIGATE PRIVACY-FRIENDLY SSP TECHNOLOGY

Further research should focus on Supply Side Platforms, in particular on privacy friendly SSP header bidding. For instance, one may encourage research in SSP bidding strategies for privacy-friendly *contextual* advertising in contrast to privacy-intrusive *(re)targeting* business models. Such research could improve RTB bidding strategies by taking notice of expressions of an end-user (1) exercising his right to object, (2) granting consent, or (3) revoking his consent.

We already know that (a) the end-user, (b) the publisher, (c) the SSP, and (d) the advertisers form a network. The publisher is in a unique position due to the direct connection to an end-user and the SSP. The SSP is in a unique position due to the direct connection to a publisher and the advertisers. We mention two possibilities.

First, special attention could be given to embedding the end-user's granular consent expressions by using technical settings of a browser square in the center of the design of SSP header-bidding protocols. Second, special attention could be given to (further) exploration of the combination of granular consent expressions with centralized tag-management solutions. For instance, firing a pixel can be made conditional on the W3C DNT settings of a browser (cf. Fed [2018, pp. 1–2]). Such use of Tag-Based Integration is, already today, available for publishers and SSPs.

In closing, we remark that the publisher can increase competition by connecting to multiple SSPs (cf. Jauvion et al. [2018]). Together they can guard the rights and freedoms of the end-user, among others, by seeking consent to be relied upon by multiple controllers (as defined in Article 4(7) GDPR).

References

NOTE: In the event that online references disappeared over time, they may still be accessible via Internet Archive, URL: <https://archive.org/>.

- Abse, N. (2012). *IAB Presents Innovations In Web Marketing and Advertising: Big data delivers on campaign promise, Microtargeted political advertising in Election 2012*. IABureau.
- Acar, G., Eubank, C., Englehardt, S., Juarez, M., Narayanan, A., & Diaz, C. (2014). The Web never forgets: Persistent tracking mechanisms in the wild. In *Proceedings of the 2014 ACM SIGSAC Conference on Computer and Communications Security* (pp. 674–689).
- Acar, G., Juarez, M., Nikiforakis, N., Diaz, C., Gürses, S., Piessens, F., & Preneel, B. (2013). FPDetective: dusting the Web for fingerprinters. In *Proceedings of the 2013 ACM SIGSAC Conference on Computer & Communications Security* (pp. 1129–1140).
- Adams, C., Fedor, F., Felten, C., Felten, E., Fenton, J., Mayer, J., Ralli, C., Toubiana, V., Turner, D., & Van Eijk, R. (2012, April 14–15). *Tracking Report Card*. [Firefox add-on]. New York, NY: Wall Street Journal. WSJ Data Transparency Code-a-Thon. Retrieved 13 April 2014, from <https://www.hackerleague.org/hackathons/wsj-data-transparency-code-a-thon/hacks/tracking-report-card>
- Alabbas, A., & Bell, J. (2018). *Indexed Database API 2.0*. W3C Recommendation, 30 January 2018. Retrieved 31 January 2018, from <https://www.w3.org/TR/IndexedDB-2/>
- Aleem, M., Ishtiaq, A., Abbasi, S., & Islam, M. (2017). User tracking mechanisms and counter measures. *International Journal of Applied Mathematics, Electronics and Computers*, 5(2), 33–40.
- Alexa (n.d.). *Top 1,000,000 sites*. [ZIP-archive, updated daily]. Alexa Internet Inc. Retrieved from <http://s3.amazonaws.com/alexa-static/top-1m.csv.zip>
- Alexa (2015a, June 20). *How are Alexa's traffic rankings determined?* [FAQ]. Alexa Internet Inc. Retrieved 20 June 2015, from

- <https://support.alexa.com/hc/en-us/articles/200449744-How-are-Alexa-s-traffic-rankings-determined->
- Alexa (2015b, June 20). *What's going on with my Alexa Rank?* [FAQ]. Alexa Internet Inc. Retrieved 20 June 2015, from <https://support.alexa.com/hc/en-us/articles/200449614>
- Al-Fannah, N., Li, W., & Mitchell, C. (2018). Beyond Cookie Monster Amnesia: Real World Persistent Online Tracking. , 21.
- Andersdotter, A., Castex, F., Dubois, N., Hamilton, S., Madelin, R., & Van Eijk, R. (2014, January 21). *Do Not Track, is self-regulation enough?* [Video]. Brussels: Eurpean Parliament. Retrieved 27 January 2014, from http://greenmediabox.eu/archive/2014/01/21/do-not-track/20140121_do-not-track_or.mp4
- Andersdotter, A., & Jensen-Urstad, A. (2016). Evaluating Websites and Their Adherence to Data Protection Principles: Tools and Experiences. In *IFIP International Summer School on Privacy and Identity Management* (pp. 39–51).
- Anderson, C. (2007). *The long tail: how endless choice is creating unlimited demand*. Random House.
- Apple (2005). The WebKit Open Source Project [Computer software manual]. Retrieved 25 June 2015, from <https://www.webkit.org/>
- AppNexus (2017a). *How to Use a Deal ID on AppNexus*. [Documentation]. Retrieved 29 January 2018, from <https://support.sharethrough.com/hc/en-us/articles/115001749003-How-to-Use-a-Deal-ID-on-AppNexus>
- AppNexus (2017b). *Online Advertising and Ad Tech Glossary*. [Wiki]. Retrieved 26 November 2017, from <https://wiki.appnexus.com/display/industry/Online+Advertising+and+Ad+Tech+Glossary>
- AppNexus (2017c). *Ways of Buying Ad Inventory*. [Wiki]. Retrieved 26 November 2017, from <https://wiki.appnexus.com/display/industry/Ways+of+Buying+Ad+Inventory>
- Article 29 Working Party (2010). *Opinion 2/2010 on online behavioural advertising*. (WP 171). Brussels: Article 29 Working Party.
- Article 29 Working Party (2011). *Opinion 13/2011 on Geolocation services on smart mobile device*. (WP 185). Brussels: Article 29 Working Party.
- Article 29 Working Party (2012). *Opinion 04/2012 on Cookie Consent Exemption*. (WP 194). Brussels: Article 29 Working Party.

- Article 29 Working Party (2013a). *Opinion 03/2013 on purpose limitation*. (WP 203). Brussels: Article 29 Working Party.
- Article 29 Working Party (2013b). *Working Document 02/2013 providing guidance on obtaining consent for cookies*. (WP 208). Brussels: Article 29 Working Party.
- Article 29 Working Party (2014a). *Article 29 Data Protection Working Party comments in response to W3C's public consultation on the W3C Last Call Working Draft, 24 April 2014, Tracking Preference Expression (DNT)*. [Letter]. Brussels: Article 29 Working Party.
- Article 29 Working Party (2014b). *Opinion 05/2014 on Anonymisation Techniques*. (WP 216). Brussels: Article 29 Working Party.
- Article 29 Working Party (2014c). *Opinion 9/2014 on the application of Directive 2002/58/EC to device fingerprinting*. (WP 224). Brussels: Article 29 Working Party.
- Article 29 Working Party (2015a). *Cookie sweep combined analysis*. (WP 229). Brussels: Article 29 Working Party.
- Article 29 Working Party (2015b). *Letter from the Art. 29 WP on W3C's public consultation on the W3C Last Call Working Draft, 14 July 2015, Tracking Compliance and Scope*. [Letter]. Brussels: Article 29 Working Party.
- Article 29 Working Party (2018). *Guidelines on consent under Regulation 2016/679*. (WP 259 rev.01). Brussels: Article 29 Working Party.
- Asghari, H., Van Eijk, R., Englehardt, S., Narayanan, A., & Winter, P. (2016). *Cookies, stroopwafels, and tea biscuits: On the many national implementations of the E.U. Cookie directive*. [Working Title, draft paper].
- Audit Bureau of Circulations (2013). *Guide to ABC's Spiders & Robots Service*. Berkhamsted, Hertfordshire: ABC. Retrieved 22 July 2015, from http://www.abc.org.uk/Documents/Spiders%20and%20Robots%20service/Guide_ABC%20Spiders%20%20Robots%20Service_v9.pdf
- Audit Bureau of Circulations (2015). *Reporting Standards Website Traffic*. In *Reporting Standards (JICWEBS)*. Berkhamsted, Hertfordshire: Audit Bureau of Circulations (ABC). Retrieved 17 July 2015, from <http://www.abcstandards.org.uk/images/ABCWebTrafficReportingStandards.pdf>
- Aziz, A., & Telang, R. (2015). *What is a Cookie Worth?* In *14th annual workshop on the economics of information security WEIS2015*. [Preliminary draft].

- Balseiro, S., Feldman, J., Mirrokni, V., & Muthukrishnan, S. (2014). Yield optimization of display advertising with ad exchange. *Management Science*, 60(12), 2886–2907.
- Barabási, A. (2016). *Network science*. Cambridge university press.
- Barnes, R., Schneier, B., Jennings, C., Hardie, T., Trammell, B., Huitema, C., & Borkmann, D. (2015, August). *Confidentiality in the Face of Pervasive Surveillance: A Threat Model and Problem Statement*. Retrieved from <http://tools.ietf.org/rfc/rfc7624.txt> (RFC7624)
- Barrat, A., Barthelemy, M., & Vespignani, A. (2007). The Architecture of Complex Weighted Networks: Measurements and Models. In *Large Scale Structure And Dynamics Of Complex Networks: From Information Technology to Finance and Natural Science* (pp. 67–92). World Scientific.
- Barth, A. (2011, April). *HTTP State Management Mechanism*. Retrieved from <http://tools.ietf.org/rfc/rfc6265.txt> (RFC6265)
- Bashir, M., Arshad, S., Robertson, W., & Wilson, C. (2016). Tracing Information Flows Between Ad Exchanges Using Retargeted Ads. In *USENIX Security Symposium* (pp. 481–496).
- Bashir, M., & Wilson, C. (2018). Diffusion of User Tracking Data in the Online Advertising Ecosystem. *Proceedings on Privacy Enhancing Technologies*, 4, 85–103.
- Basques, K. (2018). Network Analysis Reference [Computer software manual]. [Chrome Developer Tools]. Retrieved 3 September 2018, from <https://developers.google.com/web/tools/chrome-devtools/network-performance/reference>
- Baudot, F., Delcroix, G., Le Grand, G., Petitcolas, S., Jagdish, P., Castelluccia, C., Lefruit, J., & Roca, V. (2012). *Mobilitics*. [Presentation]. Paris: CNIL. Retrieved 4 January 2016, from http://www.cnil.fr/fileadmin/documents/La_CNIL/publications/DEIP/Mobilitics_confPresse_09042013.pdf
- Bavelas, A. (1948). A mathematical model for group structures. *Applied anthropology*, 7(3), 16–30.
- Beeswax (2017). *Beeswax API*. [Release notes]. Retrieved 29 January 2018, from <https://docs.beeswax.com/docs/release-notes>
- Belshe, M., Peon, R., & Thomson, M. (2015, May). *Hypertext Transfer Protocol Version 2 (HTTP/2)*. Retrieved from <http://tools.ietf.org/rfc/rfc7540.txt> (RFC7540)

- Benatallah, B., Casati, F., & Toumani, F. (2004). Web service conversation modeling: A cornerstone for e-business automation. *Internet Computing, IEEE*, 8(1), 46-54. doi: <http://dx.doi.org/10.1109/MIC.2004.1260703>
- Benson, A. R., Gleich, D. F., & Leskovec, J. (2016). Higher-order organization of complex networks. *Science*, 353(6295), 163-166.
- Berners-Lee, T. (2017). *Three challenges for the web, according to its inventor*. [Blogpost]. Retrieved 26 January 2018, from <https://webfoundation.org/2017/03/web-turns-28-letter/>
- Berners-Lee, T., Fielding, R., & Frystyk, H. (1996, May). *Hypertext Transfer Protocol – HTTP/1.0*. Retrieved from <http://tools.ietf.org/rfc/rfc1945.txt> (RFC1945)
- Berners-Lee, T., Fielding, R., & Masinter, L. (2005, January). *Uniform Resource Identifier (URI): Generic Syntax*. Retrieved from <http://tools.ietf.org/rfc/rfc3986.txt> (RFC3986)
- Beyer, M., & Douglas, L. (2012). The Importance of 'Big Data': A Definition. In *Gartner report (ID Number: G00235055)*. Stamford, CT: Gartner, Inc. Retrieved 26 March 2014, from <https://www.gartner.com/it-glossary/big-data/>
- Bidelman, E. (2017). *Getting Started with Headless Chrome*. [Blogpost]. Google. Retrieved 28 January 2018, from <https://developers.google.com/web/updates/2017/04/headless-chrome>
- BidSwitch (2016). *The Integration Challenge*. [Blogpost]. IPONWEB. Retrieved 8 August 2016, from <http://www.bidswitch.com/what-is-it/>
- Blok, P. (2018, February 7). *Inleiding gegevensbescherming*. [Slides]. Utrecht: Universiteit Utrecht.
- Blondel, V. D., Guillaume, J., Lambiotte, R., & Lefebvre, E. (2008). Fast unfolding of communities in large networks. *Journal of statistical mechanics: theory and experiment*, 2008(10), P10008.
- Boda, K., Földes, Á., Gulyás, G., & Imre, S. (2011). User tracking on the web via cross-browser fingerprinting. In *Nordic Conference on Secure IT Systems* (pp. 31-46).
- Boerman, S., Kruikemeier, S., & Zuiderveen Borgesius, F. (2017). Online behavioral advertising: A literature review and research agenda. *Journal of Advertising*, 46(3), 363-376.
- Bojinov, H., Michalevsky, Y., Nakibly, G., & Boneh, D. (2014). Mobile device identification via sensor fingerprinting. *arXiv preprint arXiv:1408.1416*.

- Bonacich, P. (1987). Power and centrality: A family of measures. *American journal of sociology*, 92(5), 1170–1182.
- Boogert, E. (2011). *Browsermakers gaan privacy consument beschermen*. Amsterdam: Emerce.nl. Retrieved 10 January 2014, from <http://www.emerce.nl/nieuws/browsermakers-gaan-privacy-consument-beschermen>
- Brandes, U. (2001). A faster algorithm for betweenness centrality. *Journal of mathematical sociology*, 25(2), 163–177.
- Bray, T. (2014, March). *The JavaScript Object Notation (JSON) Data Interchange Format*. Retrieved from <http://tools.ietf.org/rfc/rfc7159.txt> (RFC7159)
- Brinker, S. (2014). *Marketing Technology Landscape Super-graphic*. [Blogpost]. Chiefmartec.com. Retrieved 2 January 2015, from <http://chiefmartec.com/2014/01/marketing-technology-landscape-supergraphic-2014/>
- Brock, J. (2012). *TrackerBlock*. [Browser extension]. PrivacyChoice.org. Retrieved 21 June 2015, from <https://addons.mozilla.org/nl/firefox/addon/trackerblock/?src=userprofile>
- Brookman, J. (2015). *What is cross-device tracking*. [Presentation]. Washington, DC: Federal Trade Commission. Retrieved 29 December 2015, from https://www.ftc.gov/system/files/documents/public_events/630761/cross-device_tracking_workshop_deck.pptx
- Brookman, J., Rouge, P., Alva, A., & Yeung, C. (2017). Cross-device tracking: Measurement and disclosures. *Proceedings on Privacy Enhancing Technologies*, 2017(2), 133–148.
- Budak, C., Goel, S., Rao, J., & Zervas, G. (2014). Do-Not-Track and the Economics of Third-Party Advertising. *Boston U. School of Management Research Paper*(2505643).
- Bujlow, T., Carela-Español, V., Solé-Pareta, J., & Barlet-Ros, P. (2015). Web Tracking: Mechanisms, Implications, and Defenses. *arXiv preprint arXiv:1507.07872*.
- Bujlow, T., Carela-Español, V., Solé-Pareta, J., & Barlet-Ros, P. (2017). A survey on web tracking: Mechanisms, implications, and defenses. *Proceedings of the IEEE*, 105(8), 1476–1510.
- Burnik, J. (2017, April 14). *Fw: FOI request with regards to the use of Youtube no-cookie solution in Slovenia*. [Personal correspondence].
- Calo, R. (2014). Digital Market Manipulation. *Geo. Wash. L. Rev.*, 82, 995–1304.

- Cao, Y., Li, S., & Wijmans, E. (2017). browser fingerprinting via OS and hardware level features. In *Proceedings of Network & Distributed System Security Symposium (NDSS)*.
- Causata (2010). *The challange of identity association*. [Whitepaper]. Causata. Retrieved 29 September 2015, from https://web.archive.org/web/20110908003213/http://www.causata.com/res/pdf/Causata_Identity_Association.pdf
- Chaabane, A., Ding, Y., Dey, R., Kaafar, M., & Ross, K. (2014). A Closer Look at Third-Party OSN Applications: Are They Leaking Your Personal Information? In *International Conference on Passive and Active Network Measurement* (pp. 235–246).
- Chen, B., Yuan, S., & Wang, J. (2014). A dynamic pricing model for unifying programmatic guarantee and real-time bidding in display advertising. In *Proceedings of the Eighth International Workshop on Data Mining for Online Advertising* (pp. 1–9).
- Chen, Y. (2017). *Programmatic advertising is preparing for the first-price auction era*. [Blogpost]. Retrieved 8 July 2018, from <https://digiday.com/marketing/programmatic-advertising-readying-first-price-auction-era/>
- Chester, J. (2018, January 24). *Fwd: Xaxis Lightning Bolt Tag - Frequently Asked Questions*. [Personal correspondence].
- Chester, J., & Montgomery, K. (2017). The role of digital marketing in political campaigns. In *Internet policy review, Journal on internal regulation* (Vol. 6, p. 20). Berlin, Germany: Alexander von Humboldt Institute for internet and society. doi: 10.14763/2017.4.773
- Christl, W., & Spiekermann, S. (2016). *Networks of control: a report on corporate surveillance, digital tracking, big data & privacy*. Wien: Facultas.
- Citro, C. (2014). *Bigquery*. [Github]. Retrieved 28 June 2015, from <https://github.com/hadley/bigquery>
- Clark, N. (2015). Internet Governance: Is it finally time to drop the training wheels? In *Australian Journal of Telecommunications and the Digital Economy* (Vol. 3). doi: <http://doi.org/10.18080/ajtde.v3n1.3>
- Claypotch, M. (2017). *Using Headless Mode in Firefox*. [Blogpost]. Mozilla Hacks Blog. Retrieved 28 January 2018, from <https://hacks.mozilla.org/2017/12/using-headless-mode-in-firefox>

- Clift, F. (2014). *Chrome remote shell*. [Github]. Retrieved 23 June 2015, from https://github.com/minektur/chrome_remote_shell
- CM Summit, & BattelleMedia (2013). *Behind the banner. A visualization of the adtech ecosystem*. [WebGL visualization]. Retrieved 28 December 2013, from <http://cmsummit.com/behindthebanner/>
- Cofone, I. (2015). *Privacy tradeoffs in information technology law* (Unpublished doctoral dissertation). Erasmus Universiteit Rotterdam.
- Cofone, I. (2017). The way the cookie crumbles: online tracking meets behavioural economics. *International Journal of Law and Information Technology*, 25(1), 38–62.
- Cohn, B. S., & Marriott, M. (1958). Networks and centres of integration in indian civilization. *Journal of social Research*, 1(1), 1–9.
- Cooper, A., Tschofenig, H., Aboba, B., Peterson, J., Morris, J., Hansen, M., & Smith, R. (2013, July). *Privacy Considerations for Internet Protocols*. Retrieved from <http://tools.ietf.org/rfc/rfc6973.txt> (RFC6973)
- Cortesi, A. (2011). *Netograph*. Retrieved 5 August 2016, from <http://www.netograph.com>
- Cortesi, A. (2017). *Netograph*. Retrieved 18 September 2017, from <https://netograph.io/docs/api/overview>
- Cortesi, A., Hils, M., Kriechbaumer, T., & contributors. (2010). *Mitmproxy: A free and open source interactive HTTPS proxy*. Retrieved from <https://mitmproxy.org/> ([Version 4.0])
- Cortesi, A., Shaver, J., & Wingert, J. (2014). *har_extractor.py*. [Github]. Retrieved 19 June 2015, from https://github.com/mitmproxy/mitmproxy/blob/master/examples/har_extractor.py
- Council of Europe (1980). *Convention for the Protection of Individuals with regard to Automatic Processing of Personal Data*.
- Council of Europe (2010). *The protection of individuals with regard to automatic processing of personal data in the context of profiling. Recommendation CM/Rec(2010)13 and explanatory memorandum*. Retrieved 25 July 2016, from [https://www.coe.int/t/dghl/standardsetting/cdcj/CDCJ%20Recommendations/CMRec\(2010\)13E_Profiling.pdf](https://www.coe.int/t/dghl/standardsetting/cdcj/CDCJ%20Recommendations/CMRec(2010)13E_Profiling.pdf)

- Council of Europe (2018). *Protocol (CETS No. 223) amending the Convention for the Protection of Individuals with regard to Automatic Processing of Personal Data (ETS No. 108)*.
- Crocker, D., & Overell, P. (2008, January). *Augmented BNF for Syntax Specifications: ABNF*. Retrieved from <http://tools.ietf.org/rfc/rfc5234.txt> (RFC5234)
- Crunchbase (2007). *The Rubicon Project*. [Web service]. Retrieved 3 September 2018, from <https://www.crunchbase.com/organization/rubiconproject>
- Csárdi, G., & Nepusz, T. (2006). The igraph software package for complex network research. *InterJournal, Complex Systems*, 1695. Retrieved from <http://igraph.org>
- Csárdi, G., & Nepusz, T. (2014). *iGraph R Package [Computer software manual]*. Vienna, Austria. Retrieved 16 June 2015, from <http://igraph.org/r/#rmanual>
- Daigle, L. (2004, September). *WHOIS Protocol Specification*. Retrieved from <http://tools.ietf.org/rfc/rfc3912.txt> (RFC3912)
- Das, A., Acar, G., Borisov, N., & Pradeep, A. (2018). The Web's Sixth Sense: A Study of Scripts Accessing Smartphone Sensors.
- Das, A., Borisov, N., & Caesar, M. (2016). Tracking Mobile Web Users Through Motion Sensors: Attacks and Defenses. In *NDSS*.
- Das, A., Borisov, N., & Chou, E. (2018). Every Move You Make: Exploring Practical Issues in Smartphone Motion Sensor Fingerprinting and Countermeasures. *Proceedings on Privacy Enhancing Technologies*, 2018(1), 88–108.
- Davis, K. (2012). *Ethics of Big Data: Balancing risk and innovation*. O'Reilly Media, Inc.
- Davis, W. (2013). New Stanford Initiative Helps Browser Developers Refine Cookie-Blocking. In *Mediapost*. New York, NY: Mediapost. Retrieved 8 August 2013, from <http://www.mediapost.com/publications/article/202869/cookie-clearinghouse-project-could-block-third-par.html>
- Degeling, M., & Nierhoff, J. (2018). Tracking and Tricking a Profiler. In *Workshop on Privacy in the Electronic Society (WPES) 2018*.
- Degeling, M., Utz, C., Lentzsch, C., Hosseini, H., Schaub, F., & Holz, T. (2018). We Value Your Privacy... Now Take Some Cookies: Measuring the GDPR's Impact on Web Privacy. *arXiv preprint arXiv:1808.05096*.

- Dey, S., Roy, N., Xu, W., Choudhury, R., & Nelakuditi, S. (2014). AccelPrint: Imperfections of Accelerometers Make Smartphones Trackable. In *NDSS*.
- DLA Piper (2016). *EU law on cookies*. [Table]. Retrieved 6 July 2018, from https://www.dlapiper.com/~media/Files/Other/EU_Cookies_Update.pdf
- Doty, N. (2015). *Fingerprinting Guidance for Web Specification Authors*. [Unofficial Draft 13 November 2015]. World Wide Web Consortium (W3C). Retrieved 30 December 2015, from <https://w3c.github.io/fingerprinting-guidance/>
- Doty, N. (2017). *Fingerprinting Guidance for Web Specification Authors*. [W3C Editor's Draft 11 May 2017]. World Wide Web Consortium (W3C). Retrieved 8 August 2016, from <https://w3c.github.io/fingerprinting-guidance/>
- Doty, N., & Mulligan, D. (2013). Internet Multistakeholder Processes and Techno-Policy Standards: Initial Reflections on Privacy at the World Wide Web Consortium. *J. on Telecomm. & High Tech. L.*, 11, 135.
- Doty, N., West, H., Brookman, J., Harvey, S., & Newland, E. (2016). *Tracking Compliance and Scope (DNT)*. W3C Candidate Recommendation 26 April 2016. Retrieved 4 September 2017, from <http://www.w3.org/2011/tracking-protection/drafts/tracking-compliance.html>
- Dubbeld, L. (2004). *The regulation of the observing gaze: privacy implications of camera surveillance*. Twente University.
- Duggan, B. (2014, April 4). *Programmatic Buying: Efficiency, Effectiveness But also Lack of Knowledge & Transparency*. [Blogpost]. Retrieved 6 April 2014, from <https://www.ana.net/blogs/show/id/30025>
- Dumortier, J., & Kosta, E. (2015). *ePrivacy Directive: assessment of transposition, effectiveness and compatibility with proposed Data Protection Regulation*. [SMART 2013/007116].
- Dunn, R., Dudbridge, F., & Sanderson, C. (2005). The use of edge-betweenness clustering to investigate biological function in protein interaction networks. *BMC bioinformatics*, 6(1), 39.
- Eckersley, P. (2010a). How unique is your web browser? In *Privacy Enhancing Technologies* (pp. 1–18). Springer.
- Eckersley, P. (2010b). *A primer on information theory and privacy*. [Blogpost]. Electronic Frontier Foundation (EFF). Retrieved 1 July 2015, from <https://www.eff.org/deeplinks/2010/01/primer-information-theory-and-privacy>

- Eckersley, P. (2015, July 30). *Re: remark on DNT default + 2 questions on code*. [Personal correspondence].
- Eckersley, P., et al. (2014). *Privacy Badger: How heuristic blocking works*. [Github]. Electronic Frontier Foundation (EFF). Retrieved 30 June 2015, from <https://github.com/EFForg/privacybadgerfirefox>
- EFF (2015). *Do Not Track Compliance Policy. Discussion version [0.2 in progress]*. [Github]. Electronic Frontier Foundation (EFF). Retrieved 30 June 2015, from <https://github.com/EFForg/dnt-policy>
- Efrem, E., Svensson, J., & Neubauer, P. (2000). Neo4j open source code [Computer software manual]. [Github]. San Mateo. Retrieved 31 March June 2018, from <https://github.com/neo4j/neo4j>
- Eich, B., Bondy, B., et al. (2016). *Brave*. [Github]. San Francisco, CA: Brave Software Inc. Retrieved 4 July 2016, from <https://github.com/brave/browse-laptop>
- Ellis, B. (2014). *Real-time Analytics: Techniques to Analyze and Visualize Streaming Data*. Indianapolis, IN: John Wiley & Sons.
- Englehardt, S. (2016). *OpenWPM Crawl 100k Alexa web-sites, statefull dataset*. [Dropbox]. Retrieved 7 January 2017, from https://www.dropbox.com/s/85ul7lsyc89gdmv/2016-01_100k_stateful.tar.bz2
- Englehardt, S. (2018). *Automated discovery of privacy violations on the web*. PhD thesis. Tilburg: Princeton University.
- Englehardt, S., Acar, G., & Narayanan, A. (2017). *No boundaries: Exfiltration of personal data by session-replay scripts*. [Blogpost]. Retrieved 26 November 2017, from <https://freedom-to-tinker.com/2017/11/15/no-boundaries-exfiltration-of-personal-data-by-session-replay-scripts/>
- Englehardt, S., Eubank, C., Zimmerman, P., Reisman, D., & Narayanan, A. (2014). *Web Privacy Measurement: Scientific principles, engineering platform, and new results*. [Draft, Jun 1, 2014]. Princeton, NJ: Center for Information Technology Policy. Retrieved 17 May 2015, from <http://randomwalker.info/publications/WebPrivacyMeasurement.pdf>
- Englehardt, S., Han, J., & Narayanan, A. (2018). I never signed up for this! Privacy implications of email tracking. *Proceedings on Privacy Enhancing Technologies*, 2018(1), 109–126.
- Englehardt, S., & Narayanan, A. (2016). Online Tracking: A 1-million-site Measurement and Analysis. In *Proceedings of the*

- 2016 ACM SIGSAC Conference on Computer and Communications Security (pp. 1388–1401). New York, NY, USA: ACM. Retrieved from <http://doi.acm.org/10.1145/2976749.2978313> doi: 10.1145/2976749.2978313
- Englehardt, S., Reisman, D., Eubank, C., Zimmerman, P., Mayer, J., Narayanan, A., & Felten, E. (2015). Cookies That Give You Away: The Surveillance Implications of Web Tracking. In *Proceedings of the 24th International Conference on World Wide Web* (pp. 289–299).
- Ermakova, T., Fabian, B., Bender, B., & Klimek, K. (2018). Web Tracking-A Literature Review on the State of Research. In *Proceedings of the 51st Hawaii International Conference on System Sciences*.
- Eubank, C., Melara, M., Perez-Botero, D., & Narayanan, A. (2013). Shining the floodlights on mobile web tracking-a privacy survey. In *Proceedings of the IEEE Workshop on Web* (Vol. 2).
- European Commission (2017, January 17). *Proposal for a regulation of the European Parliament and of the Council concerning the respect for private life and the protection of personal data in electronic communications and repealing Directive 2002/58/EC (Regulation on Privacy and Electronic Communications)*. COM(2017) 10 final: 2017/0003 (COD).
- European Journalism Centre (2009). *EUfeeds, over 1,000 newspapers, updated every 20 minutes*. [Webservice]. Retrieved 3 December 2017, from <http://www.eufeds.eu/>
- Evans, C., Palmer, C., & Sleevi, R. (2015, April). *Public Key Pinning Extension for HTTP*. Retrieved from <http://tools.ietf.org/rfc/rfc7469.txt> (RFC7469)
- Evidon (2013). *Trackermap*. Retrieved 5 August 2016, from <https://www.ghostery.com/why-ghostery/trackermap/>
- Fairchild, J. (2018). *Google exchange bidding and OpenX case study: delivering increased demand and value for publishers globally*. Pasadena, CA: OpenX Technologies.
- Falissard, B. (2009). *Psy R-package version 1.0 [Computer software manual]*.
- Farrell, S., & Tschofenig, H. (2014, May). *Pervasive Monitoring Is an Attack*. Retrieved from <http://tools.ietf.org/rfc/rfc7258.txt> (RFC7258)
- Farrell, S., Wenning, R., Bos, B., Blanchet, M., & Tschofenig, H. (2015, December). *Report from the Strengthening the Internet*

- (*STRINT*) Workshop. Retrieved from <http://tools.ietf.org/rfc/rfc7687.txt> (RFC7687)
- Fed, J. (2018). *Respect Do Not Track parameter with Google Tag Manager*. [Blogpost]. Retrieved 16 July 2018, from <https://www.analyticsmania.com/post/respect-do-not-track-parameter-google-tag-manager>
- Feng, Y., Vasconcellos Vargas, D., Sakurai, K., Yu, S., & Ishikawa, T. (2017). Privacy Leakage of Job-related Information Seeking in Online Social Networks. *Information Processing Society of Japan SIG Technical Report*.
- Fieldfisher (2014). *Cookie 'consent' rule: EEA implementation*. [Table]. Retrieved 6 July 2018, from <https://www.fieldfisher.com/media/2481995/eu-cookie-consent-tracking-table.pdf>
- Fielding, R. (2015, May 27). *Re: question on HTTP header fields*. [Personal correspondence].
- Fielding, R., Gettys, J., Mogul, J., Frystyk, H., & Berners-Lee, T. (1997, January). *Hypertext Transfer Protocol – HTTP/1.1*. Retrieved from <http://tools.ietf.org/rfc/rfc2068.txt> (RFC2068)
- Fielding, R., Nottingham, M., & Reschke, J. (2014, June). *Hypertext Transfer Protocol (HTTP/1.1): Caching*. Retrieved from <http://tools.ietf.org/rfc/rfc7234.txt> (RFC7234)
- Fielding, R., & Reschke, J. (2014a, June). *Hypertext Transfer Protocol (HTTP/1.1): Message Syntax and Routing*. Retrieved from <http://tools.ietf.org/rfc/rfc7230.txt> (RFC7230)
- Fielding, R., & Reschke, J. (2014b, June). *Hypertext Transfer Protocol (HTTP/1.1): Semantics and Content*. Retrieved from <http://tools.ietf.org/rfc/rfc7231.txt> (RFC7231)
- Fielding, R., & Singer, D. (2018). *Tracking Preference Expression (DNT)*. W3C Working Group Note 20 October 2018. Retrieved 7 December 2018, from <https://w3c.github.io/dnt/drafts/tracking-dnt.html>
- Freeman, L. (1977). A set of measures of centrality based on betweenness. *Sociometry*, 35–41.
- FTC (2007). *FTC Staff Proposes Online Behavioral Advertising Privacy Principles*. FTC Staff Report. Washington, DC: Federal Trade Commission.
- FTC (2009). *FTC Staff Revises Online Behavioral Advertising Principles*. FTC Staff Report. Washington, DC: Federal Trade Commission.

- FTC (2010). *Protecting Consumers Privacy in an Era of Rapid Change*. Preliminary FTC Staff Report. Washington, DC: Federal Trade Commission.
- FTC (2013). *Speech A Rose By Any Other Name . . . Would Probably Violate the FTC Act: Shakespeare, the FTC, and Advertising*. Remarks of FTC Chairwoman Edith Ramirez Before the American Advertising Federation. Washington, DC: Federal Trade Commission.
- Funkhouser, K., Malloy, M., Alp, E., Poon, P., & Barford, P. (2018). Device Graphing by Example. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*.
- Galili, Tal, O'Callaghan, Alan, Sidi, Jonathan, Sievert, & Carson. (2017). heatmaply: an R package for creating interactive cluster heatmaps for online publishing. *Bioinformatics*. doi: 10.1093/bioinformatics/btx657
- Gallagher, R., & Greenwald, G. (2014). *How the NSA Plans to Infect "Millions" of Computers with Malware*. London: The Guardian.
- Gastineau, P. (2015). *HTTP Methods*. [Blogpost]. WhiteHat Security. Retrieved 25 July 2016, from <https://www.whitehatsec.com/blog/http-methods/>
- Geer, D. (2015). *T.S. Kuhn Revisited: Does a field make progress because it is a science, or is it a science because it makes progress?* Retrieved 8 April 2018, from <http://geer.tinho.net/geer.nsf.6i15.txt>
- Gomez, J., Pinnick, T., & Soltani, A. (2009). *KnowPrivacy*. Berkeley, CA: UC Berkeley. Retrieved 3 April 2018, from http://www.knowprivacy.org/report/KnowPrivacy_Final_Report.pdf
- Gómez-Boix, A., Laperdrix, P., & Baudry, B. (2018). Hiding in the Crowd: an Analysis of the Effectiveness of Browser Fingerprinting at Large Scale. In *WWW 2018: The 2018 Web Conference*.
- Google (n.d.-a). *Cloud Bigtable*. Retrieved 23 August 2018, from <https://cloud.google.com/bigtable/>
- Google (n.d.-b). *Developer Console Help: Advertising ID*. [FAQ]. Retrieved 2 January 2016, from <https://support.google.com/googleplay/android-developer/answer/6048248?hl=en>
- Google (n.d.-c). *How Floodlight counts conversions*. [Report]. Retrieved 26 January 2018, from <https://support.google.com/ds/answer/7305597?hl=en>
- Google (2015). *Chrome debugging protocol: Debugger protocol version 1.1* [Computer software manual]. [Chrome Devel-

- oper Tools]. Retrieved 23 June 2015, from <https://developer.chrome.com/devtools/docs/debugger-protocol>
- Google (2016a). *Ad Exchange Overview: Demand-Side Platforms*. [Video]. DoubleClick Publisher Training. Retrieved 19 February 2019, from <https://www.youtube.com/watch?v=MR9Zg1Cw8A0>
- Google (2016b). *Bidding API*. [Documentation]. Retrieved 29 January 2018, from <https://developers.google.com/ad-exchange/rtb/open-bidder/bidding-apis>
- Google (2017a). *Data Layer Helper Library*. [Github]. Google. Retrieved 18 February 2018, from <https://github.com/google/data-layer-helper>
- Google (2017b). *Google Tag Manager, web tracking developer guide: Using a Data Layer*. [Developer Guide]. Google. Retrieved 18 February 2018, from <https://developers.google.com/tag-manager/devguide>
- Google (2017c). *KLM activates data in real-time using Relay42's DMP and the Google platform*. [Blogpost]. Retrieved 21 January 2018, from <https://www.thinkwithgoogle.com/intl/en-154/insights-inspiration/case-studies/klm-activates-data-real-time-using-relay42s-dmp-and-google-platform>
- Gratton, E. (2013). *Understanding personal information: managing privacy risks*. Markham, Ont.: LexisNexis.
- Greco, A., & Shumpert, R. (2011). Real-time customer intelligence with Paul Phillips. In *Beyond Web Analytics Podcast, Episode 45*. Beyond Web Analytics! [Podcast]. Retrieved 12 January 2015, from <http://www.beyondwebanalytics.com/2011/05/07/episode-45/>
- Greenwald, G. (2013). *NSA collecting phone records of millions of Verizon customers daily*. London: the Guardian. Retrieved 8 March 2014, from <http://www.theguardian.com/world/2013/jun/06/nsa-phone-records-verizon-court-order>
- Grigas, P., Lobos, A., Wen, Z., & Lee, K. (2017). *Profit Maximization for Online Advertising Demand-Side Platforms*. [Presentation]. Halifax: KDD 2017. Retrieved 8 April 2018, from <https://drive.google.com/file/d/0BwF-hgLdpCD6ZFp4eGx0Wkt2Z1E/view>
- Grigorik, I. (2013). *HTTP Archive + BigQuery = Web Performance Answers*. [Google BigQuery]. Retrieved 28 June 2015, from <https://bigquery.cloud.google.com>
- Grigorik, I. (2015). *Preload*. W3C Editor's Draft 11 June 2015. Retrieved 19 June 2015, from <http://w3c.github.io/preload/>

- Grimmelmann, J. (2012). *The Sabotage of Do Not Track*. [Blogpost]. Retrieved 19 November 2017, from http://laboratorium.net/archive/2012/06/19/the_sabotage_of_do_not_track
- Gulyas, G. G., Some, D. F., Bielova, N., & Castelluccia, C. (2018). To Extend or not to Extend: on the Uniqueness of Browser Extensions and Web Logins. *arXiv preprint arXiv:1808.07359*.
- Gürses, S. (2010). *Multilateral privacy requirements analysis in online social network services*. PhD thesis. Heverlee: KU Leuven.
- Guye, M., Bettus, G., Bartolomei, F., & Cozzone, P. (2010). Graph theoretical analysis of structural and functional connectivity MRI in normal and pathological brain networks. *Magnetic Resonance Materials in Physics, Biology and Medicine*, 23(5-6), 409–421.
- Haga, Y., Takata, Y., Akiyama, M., & Mori, T. (2017). Building a Scalable Web Tracking Detection System: Implementation and the Empirical Study. *IEICE TRANSACTIONS on Information and Systems*, 100(8), 1663–1670.
- Hannak, A., Sapiezynski, P., Molavi Kakhki, A., Krishnamurthy, B., Lazer, D., Mislove, A., & Wilson, C. (2013). Measuring personalization of Web search. In *Proceedings of the 22nd international conference on World Wide Web* (pp. 527–538).
- Harris, D. (2013, March 4). *The history of Hadoop: From 4 nodes to the future of data*. [Interview]. Retrieved 12 April 2014, from <http://gigaom.com/2013/03/04/the-history-of-hadoop-from-4-nodes-to-the-future-of-data/>
- Hauschke, C. (2016). Third-Party-Elemente in deutschen Bibliothekswebseiten. *Informationspraxis*, 2(2).
- Hearne, S. (2008). *Request Map Generator*. [Github]. Retrieved 29 August 2018, from <https://github.com/WP0-Foundation/webpagetest>
- Hearne, S. (2013). *Request Map Generator*. [Github]. Retrieved 29 August 2018, from <http://requestmap.webperf.tools/>
- Hearne, S. (2015). *How to Find the Third-Parties on Your Site*. [Github]. Retrieved 29 August 2018, from <https://webperf.ninja/2015/find-third-party-assets/>
- Herz, T., Hamel, F., Uebernickel, F., & Brenner, W. (2010). Deriving a research agenda for the management of multisourcing relationships based on a literature review..
- Hevner, A. R., March, S. T., Park, J., & Ram, S. (2008). Design science in information systems research. *Management Information Systems Quarterly*, 28(1), 6.

- Hickson, I., Berjon, R., Faulkner, S., Leithead, T., Doyle Navara, E., O'Connor, E., & Pfeiffer, S. (2014, October). *HTML5*. 28 October 2014. W3C Recommendation. Retrieved from <http://www.w3.org/TR/html5/> (HTML5)
- Hidayat, A. (2011). *PhantomJS*. [Github]. Retrieved 23 June 2015, from <https://github.com/ariya/phantomjs>
- Hildebrandt, M. (2008). Defining profiling: a new type of knowledge? In *Profiling the European citizen* (pp. 17–45). Springer.
- Hilts, A., & Parsons, C. (2015). *Half Baked: The Opportunity to Secure Cookie-Based Identifiers from Passive Surveillance*. Available at SSRN 2640610.
- Hipp, R. (2000). SQLite [Computer software manual]. Retrieved 12 March 2018, from <https://www.sqlite.org/docs.html>
- Hodges, J., Jackson, C., & Barth, A. (2012, November). *HTTP Strict Transport Security (HSTS)*. Retrieved from <http://tools.ietf.org/rfc/rfc6797.txt> (RFC6797)
- Hoffman-Andrews, J. (2014). *Verizon Injecting Perma-Cookies to Track Mobile Customers, Bypassing Privacy Controls*. [Blog-post]. Retrieved 30 December 2015, from <https://www.eff.org/deeplinks/2014/11/verizon-x-uidh>
- Hoofnagle, C., & Good, N. (2012a). *Methods 2012*. Retrieved 27 December 2015, from <https://www.law.berkeley.edu/centers/bclt/research/privacy-at-bclt/web-privacy-census/methods/>
- Hoofnagle, C., & Good, N. (2012b). *Web Privacy Census October 2012 (Version 2.0), Literature review*. Retrieved 27 December 2015, from <https://www.law.berkeley.edu/centers/bclt/research/privacy-at-bclt/web-privacy-census/literature-review/>
- Hoofnagle, C., & Good, N. (2015). *Methods 2015*. Retrieved 27 December 2015, from <https://www.law.berkeley.edu/centers/bclt/research/privacy-at-bclt/web-privacy-census/methods-2015/>
- IABureau (2012). *Interactive Glossary*. Interactive Advertising Bureau. Retrieved 8 January 2015, from http://www.iab.net/wiki/index.php/Real-time_bidding
- IABureau (2015). *Deal ID*. [Glossary]. Retrieved 29 January 2018, from https://wiki.iab.com/index.php/Deal_ID
- IABureau (2017a). *Header bidding: a byte-sized overview*. Retrieved 1 February 2018, from <https://www.iab.com/wp-content/uploads/2015/10/VidLab-HeaderBidding-3.27.17V10.pdf>

- IABureau (2017b). *OpenRTB 3.0 Framework. Launching Secure Supply Chain Standards. Draft for Public Comment, September 2017*. Retrieved 19 November 2017, from <https://iabtechlab.com/wp-content/uploads/2017/09/OpenRTB-3.0-Draft-Framework-for-Public-Comment.pdf>
- IABureau (2017c). *Programmatic Mobile*. Interactive Advertising Bureau [White paper]. Retrieved 4 September 2017, from <https://www.iabeurope.eu/research-thought-leadership/iab-europe-white-paper-programmatic-mobile/>
- IABureau (2017d). *Standard Header Container Integration with an Ad Server*. [draft, in public review until July 28, 2017]. Retrieved 29 January 2018, from <https://iabtechlab.com/standards/header-tag/>
- IABureau Ad Effectiveness Council (2017). *Ad Effectiveness Glossary of Terms*. Interactive Advertising Bureau. Retrieved 7 September 2017, from <https://www.iabaustralia.com.au/research-and-resources/research-resources/item/12-research-and-resource/2364-digital-advertising-glossaries>
- IABureau Audio Council (2017). *Digital Audio Glossary of Terms*. Interactive Advertising Bureau. Retrieved 7 September 2017, from <https://www.iabaustralia.com.au/research-and-resources/research-resources/item/12-research-and-resource/2364-digital-advertising-glossaries>
- IABureau Mobile Advertising Council (2017). *Mobile Advertising Glossary of Terms*. Interactive Advertising Bureau. Retrieved 7 September 2017, from <https://www.iabaustralia.com.au/research-and-resources/research-resources/item/12-research-and-resource/2364-digital-advertising-glossaries>
- IABureau Video Advertising Council (2017). *Digital Video Glossary of Terms*. Interactive Advertising Bureau. Retrieved 7 September 2017, from <https://www.iabaustralia.com.au/research-and-resources/research-resources/item/12-research-and-resource/2364-digital-advertising-glossaries>
- IANA (2015a). *Message headers*. Internet Assigned Numbers Authority (IANA). Retrieved 17 June 2015, from <http://www.iana.org/assignments/message-headers/message-headers.xhtml> (Last Updated 5 June 2015)
- IANA (2015b). *Protocol Registries*. Internet Assigned Numbers Authority (IANA). Retrieved 17 June 2015, from <https://www>

- .iana.org/protocols
- ICDPPC (2013). *Resolution on Web tracking and Privacy*. Warsaw: International Conference of Data Protection and Privacy Commissioners.
- Improve Digital (2015). *NL Market Map*. [Infograph]. Author. Retrieved 12 December 2015, from http://www.improvedigital.com/main/wp-content/uploads/2015/08/MarketMap_NL_2015.pdf
- Improve Digital (2017a). *2017 Display Advertising Ecosystem Belgium*. [Infograph]. Author. Retrieved 4 September 2017, from http://www.improvedigital.com/main/wp-content/uploads/MarketMap_Belgium.pdf
- Improve Digital (2017b). *2017 Display Advertising Ecosystem France*. [Infograph]. Author. Retrieved 4 September 2017, from http://www.improvedigital.com/main/wp-content/uploads/MarketMap_France.pdf
- Improve Digital (2017c). *2017 Display Advertising Ecosystem Germany*. [Infograph]. Author. Retrieved 4 September 2017, from http://www.improvedigital.com/main/wp-content/uploads/MarketMap_German.pdf
- Improve Digital (2017d). *2017 Display Advertising Ecosystem Italy*. [Infograph]. Author. Retrieved 4 September 2017, from http://www.improvedigital.com/main/wp-content/uploads/MarketMap_Italy.pdf
- Improve Digital (2017e). *2017 Display Advertising Ecosystem Netherlands*. [Infograph]. Author. Retrieved 4 September 2017, from http://www.improvedigital.com/main/wp-content/uploads/MarketMap_Netherlands.pdf
- Improve Digital (2017f). *2017 Display Advertising Ecosystem Nordics*. [Infograph]. Author. Retrieved 4 September 2017, from http://www.improvedigital.com/main/wp-content/uploads/MarketMap_Nordics.pdf
- Improve Digital (2017g). *2017 Display Advertising Ecosystem Spain*. [Infograph]. Author. Retrieved 4 September 2017, from http://www.improvedigital.com/main/wp-content/uploads/MarketMap_Spain.pdf
- Inter-University Consortium for Political and Social Research (n.d.). *Data*. Retrieved 27 November 2018, from <https://www.icpsr.umich.edu/icpsrweb/ICPSR/support/glossary>
- Internet Privacy Workshop*. (2010, December 8-9). [Workshop report]. Boston, MA: Internet Architecture Board (IAB),

- W3C, ISOC, and MIT CSAIL. Retrieved 11 April 2014, from http://www.iab.org/wp-content/IAB-uploads/2011/03/PrivacyWorkshop_MeetingMinutes.pdf
- Iordanou, C., Smaragdakis, G., Poese, I., & Laoutaris, N. (2018). Tracing Cross Border Web Tracking. In *Proceedings of the ACM Internet Measurement Conference 2018*.
- ISO (2011). *ISO 24760-1:2011 A framework for identity management, Part 1: Terminology and concepts*. Retrieved 13 November 2017, from <https://www.iso.org/standard/57914.html>
- ISO (2013). *ISO 639-1:2002 Codes for the representation of names of languages. Part 1: Alpha-2 code*. Retrieved 1 Juny 2015, from http://www.iso.org/iso/catalogue_detail?csnumber=22109
- Ivie, G., Gunzerath, D., & Pinelli, R. (2015, July 8). *Update on Desktop Viewable Impression Reconciliation Guidance and Reconciliation Checklist*. [Email]. New York, NY: Media Rating Council, Inc.
- IWGDPT (2013). *Working Paper on Web Tracking and Privacy: Respect for context, transparency and control remain essential*. Berlin: Berliner Beauftragter für Datenschutz und Informationsfreiheit. (First official draft by R. van Eijk)
- IWGDPT (2014). *Working Paper on Web Tracking and Privacy: Respect for context, transparency and control remain essential*. In Dr. Alexander Dix (Ed.), *International Documents on Data Protection in Telecommunications and Media* (Vols. 1983–2013, pp. 492–514). Berlin: Berliner Beauftragter für Datenschutz und Informationsfreiheit. (Working paper adopted in 2013. First official draft by Van Eijk, R.)
- Jacobs, A. (2009). The pathologies of big data. *Communications of the ACM*, 52(8), 36–44. doi: <http://dx.doi.org/10.1145/1536616.1536632>
- Janc, A., & Zalewski, M. (2014). *Technical analysis of client identification mechanisms*. [The Chromium Projects, wiki page]. Retrieved 10 June 2015, from <https://www.chromium.org/Home/chromium-security/client-identification-mechanisms>
- Jaquith, M., Firas, Error, & Skeltoac. (2014). *Bad Behavior*, version 2.2.16. [Computer software manual]. Retrieved 6 July 2015, from <http://bad-behavior.ioerror.us/about/>
- Jauvion, G., Grislain, N., Dkengne, P., Garivier, A., & Gerchinovitz, S. (2018, July). Optimization of a SSP's Header Bidding Strategy using Thompson Sampling. *arXiv:1807.03299 [cs, stat]*.

- Retrieved 2018-07-14, from <http://arxiv.org/abs/1807.03299> (arXiv: 1807.03299) doi: 10.1145/3219819.3219917
- Jayaram, M., Kutterer, C., Kwasny, S., Runnegar, C., Seltzer, W., Van Eijk, R., & Zorbas, K. (2012, November 6-9). Who is following me: tracking the trackers. In *Internet Governance Forum*. Baku, Azerbaijan. [Video]. Retrieved 17 May 2015, from https://www.youtube.com/watch?v=mtPqt906u_g
- Kahle, B. (1996). *HTTP Archive*. [Github]. The Internet Archive. Retrieved 28 June 2015, from <https://github.com/HTTPArchive>
- Kamara, I., & Kosta, E. (2016). Do Not Track initiatives: regaining the lost user control. *International Data Privacy Law*, 6(4), 276–290.
- Kamphuis, B. (2013a, July 8). Handel in een fractie van een seconde. In *NOS Eight O’Clock National Evening News*. Hilversum: NOS. [Video]. Retrieved 17 May 2015, from <http://nos.nl/video/527311-handel-in-een-fractie-van-een-seconde.html>
- Kamphuis, B. (2013b). *In Hoe is real time bidding gegroeid?* Hilversum: NOS. [Video]. Retrieved 25 September 2013, from <http://nos.nl/video/527121-hoe-is-real-time-bidding-gegroeid.html>
- Kaye, D. (2015). *Report of the Special Rapporteur on the promotion and protection of the right to freedom of opinion and expression*, David Kaye. A /HRC/29/32. New York, NY: United Nations.
- Kennish, B., Chau, A., & Oppenheim, C. (2013). *Disconnect*. [Github]. Disconnect. Retrieved 4 July 2016, from <https://github.com/disconnectme/disconnect>
- Klitou, D. (2012). *Privacy-invading technologies: safeguarding privacy, liberty & security in the 21st century* (Unpublished doctoral dissertation). Centre for Law in the Information Society, Faculty of Law, Leiden University.
- Klyne, G., Nottingham, M., & Mogul, J. (2004, September). *Registration Procedures for Message Header Fields*. Retrieved from <http://tools.ietf.org/rfc/rfc3864.txt> (RFC3864)
- Kobusińska, A., Pawluczuk, K., & Brzeziński, J. (2018). Big Data fingerprinting information analytics for sustainability. *Future Generation Computer Systems*.
- Konrad, T. (2017). *Ghostery Unveils Most Advanced AI-Powered Privacy Protection Tool on the Market*. [Blogpost]. Cliqz International.

- Retrieved 12 December 2017, from <https://www.ghostery.com/press/ghostery-8-ai-powered-privacy/>
- Koot, M. (2012). *Measuring and predicting anonymity*. PhD thesis. Amsterdam: University of Amsterdam.
- Koot, M., Mandjes, M., Van't Noordende, G., & De Laat, C. (2011). Efficient probabilistic estimation of quasi-identifier uniqueness. In *Proceedings of NWO ICT.Open*. Utrecht: STW Technology Foundation.
- Kosinski, M. (2014). *Measurement and prediction of individual and group differences in the digital environment* (Unpublished doctoral dissertation). Ph. D. dissertation, Dept. Psychol., Cambridge Univ., Cambridge, UK.
- Kosinski, M., Stillwell, D., & Graepel, T. (2013). Private traits and attributes are predictable from digital records of human behavior. *Proceedings of the National Academy of Sciences*, 110(15), 5802–5805.
- Kostiainen, A. (2018). *Ambient Light Sensor*. W3C Candidate Recommendation, 20 March 2018. Retrieved 30 march 2018, from <https://www.w3.org/TR/2018/CR-ambient-light-20180320/>
- Kostiainen, A., & Bhaumik, R. (2018). *Magnetometer*. W3C Candidate Recommendation, 20 March 2018. Retrieved 30 march 2018, from <https://www.w3.org/TR/2018/CR-magnetometer-20180320/>
- Kostiainen, A., & Lamouri, M. (2014). *Battery Status API*. W3C Candidate Recommendation 09 December 2014. Retrieved 30 December 2015, from <http://www.w3.org/TR/battery-status/>
- Kostiainen, A., & Pozdnyakov, M. (2018). *Gyroscope*. W3C Candidate Recommendation, 20 March 2018. Retrieved 30 march 2018, from <https://www.w3.org/TR/2018/CR-gyroscope-20180320/>
- Kostiainen, A., & Shalamov, A. (2018). *Accelerometer*. W3C Candidate Recommendation, 20 March 2018. Retrieved 30 march 2018, from <https://www.w3.org/TR/2018/CR-accelerometer-20180320/>
- Kotler, P., & Armstrong, G. (2012). *Principles of marketing*. Pearson Education. 14th ed.
- Kranch, M., & Bonneau, J. (2015). Upgrading HTTPS in Mid-Air: An Empirical Study of Strict Transport Security and Key Pinning. In *Network and Distributed System Security (NDSS) Symposium*.

- Krishnamurthy, B. (2017, April 25). *Re: Follow up youtube-nocookie*. [Personal correspondence].
- Krishnamurthy, B., Malandrino, D., & Wills, C. (2007). Measuring privacy loss and the impact of privacy protection in web browsing. In *Proceedings of the 3rd symposium on Usable privacy and security* (pp. 52–63).
- Krishnamurthy, B., & Wills, C. (2006). Generating a privacy footprint on the internet. In *Proceedings of the 6th ACM SIGCOMM conference on Internet measurement* (pp. 65–70).
- Krishnamurthy, B., & Wills, C. (2009a). On the leakage of personally identifiable information via online social networks. In *Proceedings of the 2nd ACM workshop on Online social networks* (pp. 7–12).
- Krishnamurthy, B., & Wills, C. (2009b). Privacy diffusion on the web: a longitudinal perspective. In *Proceedings of the 18th international conference on World wide web* (pp. 541–550).
- Krishnamurthy, B., & Wills, C. (2010, March 21–26). *Privacy leakage on the Internet*. [Presentation]. Anaheim, CA: IETF 77.
- Krishnamurthy, B., & Wills, C. (2012). *WPM Privacy Measurement Bibliography follow-up*. Retrieved 27 December 2015, from <http://web.cs.wpi.edu/~cew/share/wpm.pdf>
- Kristol, D., & Montulli, L. (1997, February). *HTTP State Management Mechanism*. Retrieved from <http://tools.ietf.org/rfc/rfc2109.txt> (RFC2109)
- Kristol, D., & Montulli, L. (2000, October). *HTTP State Management Mechanism*. Retrieved from <http://tools.ietf.org/rfc/rfc2965.txt> (RFC2965)
- Kulyk, O., Hilt, A., Gerber, N., & Volkamer, M. (2018). "This Website Uses Cookies": Users' Perceptions and Reactions to the Cookie Disclaimer. Internet Society. Retrieved 2018-07-07, from https://www.ndss-symposium.org/wp-content/uploads/sites/25/2018/06/eurosec2018_12_Kulyk_paper.pdf doi: 10.14722/eurosec.2018.23012
- Kurtz, A., Gascon, H., Becker, T., Rieck, K., & Freiling, F. (2016). Fingerprinting mobile devices using personalized configurations. *Proceedings on Privacy Enhancing Technologies*, 2016(1), 4–19.
- La Rue, F. (2011). *Report of the Special Rapporteur on the promotion and protection of the right to freedom of opinion and expression, Frank La Rue*. A /HRC/23/40. New York, NY: United Nations.

- Lambiotte, R., & Kosinski, M. (2014). Tracking the digital footprints of personality. *Proceedings of the IEEE*, 102(12), 1934–1939.
- Laperdrix, P. (2017). Browser Fingerprinting: Exploring Device Diversity to Augment Authentication and Build Client-Side Countermeasures. , 169.
- Laperdrix, P., Rudametkin, W., & Baudry, B. (2015). Mitigating Browser Fingerprint Tracking: Multi-level Reconfiguration and Diversification. In *Proceedings of the 10th International Symposium on Software Engineering for Adaptive and Self-Managing Systems* (pp. 98–108). Piscataway, NJ, USA: IEEE Press. Retrieved from <http://dl.acm.org/citation.cfm?id=2821357.2821378>
- Laperdrix, P., Rudametkin, W., & Baudry, B. (2016). Beauty and the Beast: Diverting modern web browsers to build unique browser fingerprints. In *37th IEEE Symposium on Security and Privacy (S&P 2016)*. San Jose, CA. (Available at HAL: hal-01285470v2)
- Laszlo, J., & Smith, B. (2015). Mobile Programmatic Playbook. *Interactive Advertising Bureau Report*. Retrieved 24 December 2015, from <http://www.iab.com/wp-content/uploads/2015/05/MobileProgrammaticPlaybook.pdf>
- Latour, B. (2007). Beware, your imagination leaves digital traces. In *Times Higher Literary Supplement*. London: News UK. Retrieved 12 August 2013, from <http://www.bruno-latour.fr/node/245>
- Lavrenovs, A., & Melon, F. (2018). HTTP Security Headers Analysis of Top One Million Websites. In T. Minárik, R. Jakschis, & L. L. (Eds.), *10th International Conference on Cyber Conflict (CyCon X): Maximising Effects*. Tallin: NATO CCD COE Publications.
- Le Hors, A., Le Hégarret, P., Wood, L., Nicol, G., Robie, J., Champion, M., & Byrne, S. (2004). *Document Object Model (DOM) Level 3 Core Specification*. Version 1.0, W3C Recommendation 07 April 2004. Retrieved 22 June 2015, from <http://www.w3.org/TR/DOM-Level-3-Core/> (DOM 3 CORE)
- Leenes, R., & Kosta, E. (2015). Taming the cookie monster with dutch law—a tale of regulatory failure. *Computer Law & Security Review*, 31(3), 317–335.
- Leung, C., Ren, J., Choffnes, D., & Wilson, C. (2016). Should you use the app for that?: Comparing the privacy implications of app-and web-based online services. In *Proceedings of the 2016 Internet Measurement Conference* (pp. 365–372).

- Levin, C., & McCain, J. (2014). *Online advertising and hidden hazards to consumer security and data privacy*. Washington, DC: United States Senate Permanent Subcommittee on Investigations, Committee on Homeland Security and Governmental Affairs.
- Libbenga, J. (2013). *Mozilla wil cookiegedrag inperken*. Amsterdam: Emerce.nl. Retrieved 8 August 2013, from <http://www.emerce.nl/nieuws/mozilla-cookiegedrag-inperken>
- Libbenga, J. (2014). *Machtsstrijd: de wereld na de cookie*. Amsterdam: Emerce.nl. Retrieved 10 Januari 2016, from <http://www.emerce.nl/nieuws/machtsstrijd-wereld-na-cookie>
- Libert, T. (2015a). Exposing the hidden web: An analysis of third-party http requests on 1 million websites. *International Journal of Communication*, 9, 3544–3561.
- Libert, T. (2015b). Privacy implications of health information seeking on the web. *Communications of the ACM*, 58(3), 68–77.
- Libert, T. (2015c). *WebXray*. [Github]. Retrieved 30 December 2015, from <https://github.com/timlib/webXray.git>
- Libert, T. (2018). An Automated Approach to Auditing Disclosure of Third-Party Data Collection in Website Privacy Policies. In *The Web Conference (WWW2018)*.
- Libert, T., & Graves, L. (2018). Third-party web content on EU news sites. Potential challenges and paths to privacy improvement. [Factsheet].
- Libert, T., Graves, L., & Nielsen, R. K. (2018). Changes in third-party content on European news websites after GDPR. [Factsheet].
- Lindskow, K. (2016). *Exploring Digital News Publishing Business Models: A Production Network Approach*. PhD thesis. Copenhagen: Copenhagen Business School.
- Loukides, M. (2010). *What is data science? The future belongs to the companies and people that turn data into products*. Sebastopol, CA: O'Reilly Radar. Retrieved 26 March 2014, from <http://radar.oreilly.com/2010/06/what-is-data-science.html>
- Lumapartners (2014). *Marketing Technology LUMAscape*. [Blogpost]. Lumapartners.com. Retrieved 2 January 2015, from <http://www.lumapartners.com/lumascape/marketing-technology-lumascape/>
- Lynch, C. (2008). Big data: How do your data grow? *Nature*, 455(7209), 28–29.
- Maass, M., Wichmann, P., Pridöhl, H., & Herrmann, D. (2017). PrivacyScore: Improving Privacy and Security via Crowd-Sourced

- Benchmarks of Websites. In *Annual Privacy Forum* (pp. 178–191).
- Majumdar, S., Kulkarni, D., & Ravishankar, C. (2007). Addressing click fraud in content delivery systems. In *INFOCOM 2007. 26th IEEE International Conference on Computer Communications*. IEEE (pp. 240–248).
- Malloy, M., Barford, P., Alp, E., Koller, J., & Jewell, A. (2017). Internet Device Graphs. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 1913–1921).
- Marquis-Boire, M., Greenwald, G., & Lee, M. (2015). *XKEYSCORE: NSA's Google for the World's Private Communications*. The Intercept. Retrieved 2 July 2015, from <https://firstlook.org/theintercept/2015/07/01/nsas-google-worlds-private-communications/>
- Mayer, J. (2015a). Surveillance Law: A Conversation with Peter Swire, Member of the President's Intelligence Review Group. [Coursera]. Retrieved 27 December 2015, from <https://class.coursera.org/surveillance-001/lecture/95>
- Mayer, J. (2015b). *The Turn-Verizon Zombie Cookie*. [Blogpost]. Retrieved 30 December 2015, from <http://webpolicy.org/2015/01/14/turn-verizon-zombie-cookie/>
- Mayer, J., & Mitchell, J. (2012). Third-party Web tracking: Policy and technology. In *Security and Privacy (SP), 2012 IEEE Symposium on* (pp. 413–427). Washington, DC: IEEE.
- Mayer, J., Narayanan, A., & Stamm, S. (2011, September). *Do Not Track: A Universal Third-Party Web Tracking Opt Out*. draft-mayer-do-not-track-00. IETF. Retrieved 11 April 2014, from <http://tools.ietf.org/id/draft-mayer-do-not-track-00.txt>
- Mayer-Schönberger, V., & Cukier, K. (2013). *Big data: A revolution that will transform how we live, work, and think*. Boston: Houghton Mifflin Harcourt.
- Mazel, J., Garnier, R., & Fukuda, K. (2017). A comparison of web privacy protection techniques. *arXiv preprint arXiv:1712.06850*.
- McDonald, A. (2018). Stakeholders and High Stakes Divergent Standards for Do Not Track. In J. Polonetsky, O. Tene, & E. Selinger (Eds.), *Cambridge Handbook of Consumer Privacy*. Cambridge: Cambridge University Press.

- McDonald, A., & Van Eijk, R. (2012, May 31-June 1). Problem identification. In *Workshop Web Privacy Measurement*. Berkeley, CA: Berkeley Center for Law & Technology.
- Media Impact (2018). *Integration von Google (DfP/AdX) in einen AppNexus AdStack*. [Whitepaper]. Media Impact. Retrieved 27 May 2018, from https://omr.mediaimpact.de/assets/img/Whitepaper_Media-Impact_AdStack.pdf
- Meesters, P. (2014). *Intelligent Blauw, Intelligence-gestuurde politiezorg in gebiedsgebonden eenheden*. PhD thesis. Tilburg: Tilburg University.
- Melnik, S., Gubarev, A., Long, J., Romer, G., Shivakumar, S., Tolton, M., & Vassilakis, T. (2010). Dremel: Interactive Analysis of Web-Scale Datasets. In *Proc. of the 36th Very Large Data Bases (VLDB) Conference* (p. 330-339). VLDB Endowment Inc.
- Meron, S., & Huh, J. (2014). *Yahoo to Acquire BrightRoll. Creates the largest digital video advertising platform*. [Blogpost]. Retrieved 9 April 2018, from <https://www.businesswire.com/news/home/20141111006545/en/Yahoo-Acquire-BrightRoll>
- Merriam-Webster.com (n.d.). *Publisher*. Retrieved 15 January 2018, from <https://www.merriam-webster.com/dictionary/publisher>
- Merzdovnik, G., Huber, M., Buhov, D., Nikiforakis, N., Neuner, S., Schmiedecker, M., & Weippl, E. (2017). Block me if you can: A large-scale study of tracker-blocking tools. In *Security and Privacy (EuroS&P), 2017 IEEE European Symposium on* (pp. 319-333).
- Meyer, S. (2017a). *EU Privacy Regulations: A Greek Tragedy*. [Blogpost]. Retrieved 4 February 2018, from <https://www.evidon.com/blog/privacy-compliance/greek-tragedy-eu-privacy-regulations/>
- Meyer, S. (2017b). *The Next Chapter: Evidon Joins Crownpeak*. [Blogpost]. Retrieved 4 February 2018, from <https://www.evidon.com/blog/evidon/evidon-joins-crownpeak/>
- MIT Technology Review Custom (2013). *The evolution of ad tech. Going from Mad Men to Math Men. How technology has fundamentally changed the art and science of advertising*. [Blogpost]. MIT Technology Review.
- Mitchell, J. (2016, March 29). *Re: conection*. [Personal correspondence].
- Moeller, B., & Langley, A. (2015, April). *TLS Fallback Signaling Cipher Suite Value (SCSV) for Preventing Protocol Downgrade*

- Attacks*. Retrieved from <http://tools.ietf.org/rfc/rfc7507.txt> (RFC7507)
- Mozilla (2011). *Lightbeam*. [Browser extension]. Retrieved 2 January 2016, from <https://www.mozilla.org/en-US/lightbeam/>
- Mozilla (2018a). *Browse without baggage in Firefox: Set Tracking Protection to always on*. [Blogpost]. Retrieved 1 February 2018, from <https://blog.mozilla.org/firefox/tracking-protection-always-on>
- Mozilla (2018b). *Privacy Possum*. [Firefox add-on]. Retrieved 27 May 2018, from <https://github.com/cowlicks/privacypossum>
- Murray, R., Caulier-Grice, J., & Mulgan, G. (2010). *The open book of social innovation*. London: National Endowment for Science, Technology and the Art.
- Narayanan, A. (2018, March 30). *Re: list of openwpm plugins?* [Personal correspondence].
- Narayanan, A., & Reisman, D. (2017). The princeton web transparency and accountability project. In *Transparent Data Mining for Big and Small Data* (pp. 45–67). Springer.
- Nassri, A. (2015). *HAR Resources: A community curated list of resources, tools, projects and applications that support HTTP Archive (HAR)*. [Github]. Retrieved 28 June 2015, from <https://github.com/ahmadnassri/har>
- Needham, K. (2015). *The Future of Developing Firefox Add-ons*. [Blogpost]. Mozilla Add-ons Blog. Retrieved 19 December 2015, from <https://blog.mozilla.org/addons/2015/08/21/the-future-of-developing-firefox-add-ons/>
- Newman, M., & Girvan, M. (2004). Finding and evaluating community structure in networks. *Physical review E*, 69(2), 026113.
- Nielsen, P., DeMille, V., Kilrain, J., Meyer, S., Donohoo, J., Kozek, E., Van Oss, B., & Signanini, J. (2016, December 15). *Methods, apparatus, and systems for surveillance of third-party digital technology vendors providing secondary content in a content publisher's web page*. Google Patents. (US Patent App. 15/059,296)
- Nikiforakis, N., Joosen, W., & Livshits, B. (2015). Privaricator: Deceiving fingerprinters with little white lies. In *Proceedings of the 24th International Conference on World Wide Web* (pp. 820–830).
- Nikiforakis, N., Kapravelos, A., Joosen, W., Kruegel, C., Piessens, F., & Vigna, G. (2013). Cookieless monster: Exploring the

- ecosystem of web-based device fingerprinting. In *IEEE Symposium on Security and Privacy*. Washington, DC: IEEE.
- Nottingham, M. (2015). *Unsanctioned Web Tracking* (W3C TAG Finding 17 July 2015). [Github]. Retrieved 19 November 2017, from <https://www.w3.org/2001/tag/doc/unsanctioned-tracking>
- NTIA (2010). *Commercial Data Privacy and Innovation in the Internet Economy: A Dynamic Policy Framework*. IPTF Privacy Green Paper. Washington, DC: National Telecommunications and Information Administration (NTIA).
- NWO (2006). *Annotated graph mining*. Available from NWO online database, OND1315812. Nederlandse Organisatie voor Wetenschappelijk Onderzoek, (NWO). Retrieved 3 January 2014, from <http://www.narcis.nl/research/RecordID/OND1315812/Language/en>
- Odvarko, J., Jain, A., & Davies, A. (2012). *HTTP Archive (HAR) format*. W3C Editor's Draft August 14, 2012. Retrieved 19 June 2015, from <https://dvcs.w3.org/hg/webperf/raw-file/tip/specs/HAR/Overview.html>
- OECD (1980). *Guidelines on the Protection of Privacy and Transborder Flows of Personal Data*. Paris: Organisation for Economic Co-operation and Development (OECD) Publishing.
- OECD (2013a). Exploring Data-Driven Innovation as a New Source of Growth: Mapping the Policy Issues Raised by Big Data. In *OECD Digital Economy Papers* (Vol. 222). Paris: Organisation for Economic Co-operation and Development (OECD) Publishing. doi: <http://dx.doi.org/10.1787/5k47zw3fcp43-en>
- OECD (2013b). *Revised Guidelines on the Protection of Privacy and Transborder Flows of Personal Data*. Paris: Organisation for Economic Co-operation and Development (OECD) Publishing.
- Ofcom (2015). *Adults' media use and attitudes*. Report 2015. [Report]. Ofcom. Retrieved 5 August 2016, from http://stakeholders.ofcom.org.uk/binaries/research/media-literacy/media-lit-10years/2015_Adults_media_use_and_attitudes_report.pdf
- Olejnik, L., Acar, G., Castelluccia, C., & Diaz, C. (2015). *The leaking battery. A privacy analysis of the HTML5 Battery Status API*. Leuven: International Association for Cryptologic Research (IACR). Retrieved 8 July 2015, from eprint.iacr.org/2015/616.pdf
- Olejnik, L., & Castelluccia, C. (2016). *Analysis of openx-publishers cooperation*.

- O'Neill, M. (2015). *TrackerScope*. Oxford: Baycloud Systems Ltd. Retrieved 2 January 2016, from <https://baycloud.com/>
- O'Neill, M. (2017, December 1). *RE: native speaker question*. [Personal correspondence].
- Oremus, W. (2017). *The Pivot From Advertising. It's time for online media to stop chasing clicks (and Facebook's carrots)*. [Blogpost]. Retrieved 29 January 2018, from <http://www.slate.com/technology/2018/01/how-facebook-s-plan-to-rate-the-media-might-actually-work.html>
- Outsource Ad Ops (2014). *The jobs of ad ops*. [Infographic]. New York, NY: Author. Retrieved 31 January 2015, from <http://visual.ly/job-ad-ops>
- Outsource Ad Ops (2015). *Managing monetization*. [Infographic]. New York, NY: Author. Retrieved 31 January 2015, from <http://visual.ly/managing-monetization>
- Pannetrat, A. (2012). *Cookieminer*. [Google Code Archive]. Retrieved 30 January 2016, from <https://code.google.com/archive/p/cookie-miner/>
- Pannetrat, A. (2016, January 23). *Credits to your programming efforts*. [Personal correspondence].
- Papadopoulos, P., Kourtellis, N., & Markatos, E. (2018). Cookie Synchronization: Everything You Always Wanted to Know But Were Afraid to Ask. *arXiv preprint arXiv:1805.10505*.
- Papadopoulos, P., Rodriguez, P. R., Kourtellis, N., & Laoutaris, N. (2017). If you are not paying for it, you are the product: How much do advertisers pay to reach you? In *Proceedings of the 2017 Internet Measurement Conference* (pp. 142–156).
- Parliament of the EU and the Council (1995, October 24). *Directive 95/46/EC of the European Parliament and of the Council of 24 October 1995 on the protection of individuals with regard to the processing of personal data and on the free movement of such data*.
- Parliament of the EU and the Council (2000a, December 18). *Directive (EU) 2016/680 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data by competent authorities for the purposes of the prevention, investigation, detection or prosecution of criminal offences or the execution of criminal penalties, and on the free movement of such data, and repealing Council Framework Decision 2008/977/JHA*.
- Parliament of the EU and the Council (2000b, December 18). *Regulation (EC) No 45/2001 of the European Parliament and of the Coun-*

- cil of 18 December 2000 on the protection of individuals with regard to the processing of personal data by the Community institutions and bodies and on the free movement of such data.*
- Parliament of the EU and the Council (2002, July 12). *Directive 2002/58/EC of the European Parliament and of the Council of 12 July 2002 concerning the processing of personal data and the protection of privacy in the electronic communications sector (Directive on privacy and electronic communications).*
- Parliament of the EU and the Council (2009, November 25). *Directive 2009/136/EC of the European Parliament and of the Council of 25 November 2009 amending Directive 2002/22/EC on universal service and users' rights relating to electronic communications networks and services, Directive 2002/58/EC concerning the processing of personal data and the protection of privacy in the electronic communications sector and Regulation (EC) No 2006/2004 on cooperation between national authorities responsible for the enforcement of consumer protection laws.* Retrieved 30 December 2015, from <http://eur-lex.europa.eu/legal-content/EN/TXT/?qid=1414070545086&uri=CELEX:32009L0136>
- Parliament of the EU and the Council (2016). *Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation).*
- Patra, C. (2015). *Amazon DynamoDB: ten things you really should know.* [Blogpost]. CloudAcademy Blog. Retrieved 26 December 2016, from <http://cloudacademy.com/blog/amazon-dynamodb-ten-things/>
- PayPal (2017). *List of Third Parties (other than PayPal Customers) with Whom Personal Information May be Shared.* Retrieved 16 March 2018, from <https://www.paypal.com/ie/webapps/mpp/ua/third-parties-list-prev>
- PayPal (2018). *List of Third Parties (other than PayPal Customers) with Whom Personal Information May be Shared.* Retrieved 16 March 2018, from <https://www.paypal.com/ie/webapps/mpp/ua/third-parties-list>
- Pels, D. (2013, July 5). *Surf-gedrag niet meer te koop: Firefox legt handel in cookies aan banden. Browsers nemen het voortouw in discussie over privacy.* In *Trouw* (p. 16). Amsterdam: Trouw.
- Peon, R., & Ruellan, H. (2015, May). *HPACK: Header Compression for HTTP/2.* Retrieved from <http://tools.ietf.org/rfc/>

- rfc7541.txt (RFC7541)
- Peppet, S. (2014). *Regulating the Internet of Things: First Steps Toward Managing Discrimination, Privacy, Security & Consent. Texas Law Review, Forthcoming.*
- Perez, F. (2001). *Jupyter*. Retrieved 2 December 2017, from <https://jupyter.org/>
- Petitcolas, S. (2013). *CookieViz*. [Github]. Paris: CNIL. Retrieved 4 January 2016, from <https://github.com/LaboCNIL/CookieViz>
- Petnel, R. (2005). *EasyList*. Retrieved 27 November 2017, from <https://easylist.to/pages/about.html>
- Phillips, P. (2015). *From Croissantonomics to Evolution at Warp Speed*. [Blogpost]. LinkedIn.com. Retrieved 12 September 2018, from <https://www.linkedin.com/pulse/from-croissantonomics-evolution-warp-speed-paul-phillips/>
- Pierce, J. (2010). *Ghostery*. [Github]. Retrieved 19 December 2015, from <https://github.com/jonpierce/ghostery>
- Podesta, J., Pritzker, P., Moniz, E., Holdren, J., & Zients, J. (2014). *Big data: seizing opportunities, preserving values*. Washington, DC: Executive Office of the President.
- Poitras, L. (2012). *The Program*. [Documentary]. Praxis Films.
- Poitras, L. (2014). *Citizenfour*. [Documentary]. Praxis Films.
- Polonetsky, J. (2018, January 24). *Re: DSP's and SSP's using tag management to track clicks conversions*. [Personal correspondence].
- Powers, S. (2008). *Learning JavaScript: Add Sparkle and Life to Your Web Pages*. "O'Reilly Media, Inc."
- Pozdnyakov, M., Shalamov, A., Christiansen, K., & Kostiainen, A. (2018). *Orientation Sensor*. W3C Candidate Recommendation, 20 March 2018. Retrieved 30 March 2018, from <https://www.w3.org/TR/2018/CR-orientation-sensor-20180320/>
- R Core Team (2014). *R: A Language and Environment for Statistical Computing* [Computer software manual]. Vienna, Austria. Retrieved 16 June 2015, from <http://www.R-project.org>
- Rannenberg, K. (2017, October 26). *NIST's recent Digital Identity Guidelines missing essential Privacy by design feature*. [IPEN mailinglist].
- Rappaport, S. (2015). *The Digital Metrics Field Guide: The Definitive Reference for Brands Using the Web, Social Media, Mobile Media, Or Email*. Bis B.V., Uitgeverij (BIS Publishers).
- Raschke, P., & Küpper, A. (2018). *Uncovering Canvas Fingerprinting in Real-Time and Analyzing Its Usage for Web Tracking. In Security & Privacy at Large - GI Informatik 2018.*

- Reagan, R. (2013). *Turn Infographic: The Life of an Ad*. [Infographic]. ExchangeWire.com. Retrieved 29 January 2015, from <https://www.exchangewire.com/blog/2013/11/11/turn-infographic-the-life-of-an-ad/>
- Reisman, D., Englehardt, S., Eubank, C., Zimmerman, P., & Narayanan, A. (2014). *Cookies that give you away: Evaluating the surveillance implications of web tracking*. [Draft: April 2, 2014]. Princeton, NJ: Center for Information Technology Policy. Retrieved 30 June 2015, from <http://randomwalker.info/publications/cookie-surveillance.pdf>
- Reitz, K., Benfield, C., Cordasco, I., & Petrov, A. (2011). *Requests: HTTP for Humans*. [Github]. Retrieved 15 December 2015, from <http://docs.python-requests.org/en/latest/index.html#requests-http-for-humans>
- Relay42 (n.d.). *Relay42 tag management*. [Factsheet]. Retrieved 26 January 2018, from <https://form.relay42.com/factsheet/relay42-tag-management>
- Ren, J., Lindorfer, M., Dubois, D. J., Rao, A., Choffnes, D., & Vallina-Rodriguez, N. (2018). Bug Fixes, Improvements,... and Privacy Leaks. In *NDSS*.
- Rescorla, E., & Internet Architecture Board. (2005, June). *Writing Protocol Models*. Retrieved from <http://tools.ietf.org/rfc/rfc4101.txt> (RFC4101)
- Rich, J. (2015, January 21). *Beyond Cookies: Privacy Lessons for Online Advertising*. Washington, DC: Fedetal Trade Commission, Bureau of Consumer Protection.
- Richards, N., & King, J. (2014). Big Data Ethics. *Available at SSRN 2384174*.
- Riesen, K., Jiang, X., & Bunke, H. (2010). Exact and inexact graph matching: Methodology and applications. In C. C. Aggerwal & H. Wang (Eds.), *Managing and Mining Graph Data* (pp. 217–247). Springer.
- Robinson, I., Webber, J., & Eifrem, E. (2015). *Graph databases* (Second Edition ed.). O'Reilly Media, Inc.
- Rosenberg, J., & Jennings, C. (2008, January). *The Session Initiation Protocol (SIP) and Spam*. Retrieved from <http://tools.ietf.org/rfc/rfc5039.txt> (RFC5039)
- Rosenstein, M., Collins, J., & De Luca, C. (1993). A practical method for calculating largest Lyapunov exponents from small data sets. *Physica D: Nonlinear Phenomena*, 65(1), 117–134.

- Russell, A., Son, J., Archibald, J., & Kruisselbrink, M. (2018). *Service Workers 1*. W3C Editor's Draft, 13 January 2018. Retrieved 31 January 2018, from <https://w3c.github.io/ServiceWorker/v1>
- Russom, P. (2011). Big data analytics. *TDWI Best Practices Report, Fourth Quarter*. Retrieved 14 April 2014, from <http://public.dhe.ibm.com/common/ssi/ecm/en/iml14293usen/IML14293USEN.PDF>
- Rutgers, A. (2008). *Anonymous Subscriber ID (ASID) - A Mobile Web Site Developers Cookbook*. [Blogpost]. Open Mobiel Internet. Retrieved 30 December 2015, from <http://www.t-mobile.nl/corporate/media/pdf/asid-omi-cookbook.pdf>
- Ryan, J. (2017). *Deck at GDPR Summit at Croke Park*. [Presentation]. Dublin: pageFair. Retrieved 28 January 2018, from <https://www.slideshare.net/JohnnyRyan/deck-at-gdpr-summit-at-croke-park>
- Ryan, J. (2018). *Re: Feedback on Guidelines on Consent under Regulation 2016/679 (wp259)*. [Letter]. Retrieved 6 July 2018, from <https://www.dropbox.com/s/cnumg4rcw5j7im3/Letter%20to%20Article%2029%20Working%20Party%2C%2023%20January%202018.pdf>
- Savard, D., & Co, N. (2011). *Live HTTP Headers*. [Mozdev repository]. Retrieved 21 December 2015, from <http://www.mozdev.org/source/browse/livehttpheaders/>
- Scheepers, K. (2018). *Respect voor de meldplicht: de motieven achter het overtreden van de meldplicht datalekken bij de Autoriteit Persoonsgegevens*. Master's thesis. Leiden: Instituut voor Strafrecht en Criminologie, Leiden University.
- Scheitle, Q., Jelten, J., Hohlfeld, O., Ciprian, L., & Carle, G. (2018). Structure and Stability of Internet Top Lists. *arXiv preprint arXiv:1802.02651*.
- Schermer, B. (2007). *Software agents, surveillance, and the right to privacy : a legislative framework for agent-enabled surveillance*. PhD thesis. Leiden: Leiden University Press.
- Schneider, M., Enzmann, M., & Stopczynski, M. (2014). *Web-tracking-report 2014*. Fraunhofer-Institut für Sichere Informationstechnologie (SIT) technical report, SIT-TR-2014-01. Darmstadt: Fraunhofer Verlag.
- Schneier, B. (2013). *Metadata Equals Surveillance*. [Blogpost]. Retrieved 8 March 2014, from https://www.schneier.com/blog/archives/2013/09/metadata_equals.html

- Schoch, D. (2015). *A Positional Approach for Network Centrality* (Unpublished doctoral dissertation). Ph. D. dissertation, Universität Konstanz.
- Schütte, J. (2011). *Prividor*. Fraunhofer-Institut für Angewandte und Integrierte Sicherheit. Garching: Fraunhofer AISEC. Retrieved 15 July 2015, from <http://www.aisec.fraunhofer.de/de/fields-of-expertise/projekte/prividor.html>
- Secretary of State (2018). *An act to add Title 1.81.5 (commencing with Section 1798.100) to Part 4 of Division 3 of the Civil Code, relating to privacy (AB-375)*. Retrieved 29 June 2018, from http://leginfo.legislature.ca.gov/faces/billNavClient.xhtml?bill_id=201720180AB375.
- Selenium Project (2004). Selenium Documentation [Computer software manual]. Retrieved 17 June 2015, from <http://www.seleniumhq.org/docs/>
- Shannon, C. (1948). A mathematical theory of communication. In *Bell System Technical Journal* (Vol. 27, pp. 379–423, 623–656). American Telephone and Telegraph Co.
- Shaw, M. E. (1954). Group structure and the behavior of individuals in small groups. *The Journal of psychology*, 38(1), 139–149.
- Shields, R. (2014). *The Rise Of The Mega Stacks*. [Blog-post]. ExchangeWire.com. Retrieved 18 September 2014, from <https://www.exchangewire.com/blog/2014/09/15/the-rise-of-the-mega-stacks/>
- Shimbel, A. (1953). Structural parameters of communication networks. *The bulletin of mathematical biophysics*, 15(4), 501–507.
- SIDN (2014). *Jaarverslaggeving over 2014*. Arnhem: Stichting Internet Domeinregistratie Nederland (SIDN). Retrieved 21 June 2015, from <https://www.sidn.nl/contentAsset/raw-data/1fa00328-fb77-4daf-9c4c-7fa33b076541/fileAsset>
- Singer, D. (2015, May 27). *Re: question on HTTP header fields*. [Personal correspondence].
- Singer, N. (2012a). Anúncios em sites são vendidos em leilões no momento em que o internauta visita a página. In *Folha de S.Paulo*. São Paulo: Folha de S.Paulo.
- Singer, N. (2012b). Empresas de marketing calculan al instante las opciones de compra del internauta. Preocupación por la discriminación del público según su nivel adquisitivo. In *El País*. Madrid: El País.

- Singer, N. (2012c, November 18). You for sale: Your Attention, Bought in an Instant. In *New York Times* (p. BU1). New York, NY: New York Times.
- Smith, C. (2014). *FBX Is Turning Facebook Into One Of The Biggest Players In The Real-Time Bidding Space*. [Blog-post]. Business Insider. Retrieved 13 June 2015, from <http://www.businessinsider.com/explaining-fbx-facebook-exchange-2-2014-1?IR=T>
- Snowden, E. (2014). *Testimony to the European Parliament*. Retrieved 13 March 2014, from <http://www.europarl.europa.eu/document/activities/cont/201403/20140307ATT80674/20140307ATT80674EN.pdf>
- Soghoian, C. (2011). *The History of the Do Not Track Header*. [Blog-post]. Retrieved 4 April 2014, from <http://paranoia.dubfire.net/2011/01/>
- Soltani, A. (2014). *More evidence of the NSA piggybacking on commercial tracking by Google, Yahoo, Facebook in order to identify users*. [Twitter]. Retrieved 25 May 2015, from <https://twitter.com/ashk4n/status/443771573489913856/photo/1>
- Soltani, A., Cortesi, A., & Campbell, D. (2012). *MobileScope [Computer software manual]*. Retrieved 21 June 2015, from <http://ashkansoltani.org/work/mobile/#mobilescope>
- Some, D., Rezk, T., & Bielova, N. (2018). *Web Stats*. [Blogpost]. Retrieved 31 March 2018, from <https://webstats.inria.fr/popjslibs.php>
- Sørensen, J., & van den Bulck, H. (2018). Public Service Media Online, Advertising and the Third-Party User Data Business: A Trade versus Trust Dilemma? *Convergence. the International Journal of Research Into New Media Technologies*.
- Starov, O., & Nikiforakis, N. (2017a). Extended tracking powers: Measuring the privacy diffusion enabled by browser extensions. In *Proceedings of the 26th International Conference on World Wide Web* (pp. 1481–1490).
- Starov, O., & Nikiforakis, N. (2017b, May). XHOUND: Quantifying the Fingerprintability of Browser Extensions. In (pp. 941–956). IEEE. Retrieved 2018-04-02, from <http://ieeexplore.ieee.org/document/7958618/> doi: 10.1109/SP.2017.18
- Stewart, S., & Burns, D. (2018). *Webdriver*. W3C Proposed Recommendation 26 April 2018. Retrieved 27 May 2018, from <https://www.w3.org/TR/webdriver/>

- Swift, M. (2014, June 20). Advertising groups blast 'Do Not Track' effort, but European and US leaders push privacy plan forward. In *MLex Market Insight*. Brussels/San Francisco: MLex.
- Swire, P. (2013). *Explanatory Memorandum for Working Group Decision on "What Base Text to use for the Do Not Track Compliance Specification"*. Retrieved 27 July 2015, from <http://www.w3.org/2011/tracking-protection/2013-july-decision/>
- Swire, P. (2015). *US Surveillance Law, Safe Harbor, and Reforms since 2013*. Retrieved 18 December 2015, from <http://fpf.org/wp-content/uploads/2015/12/White-Paper-Swire-US-EU-Surveillance.pdf>
- Swire, P. (2017). *Professor Peter Swire Testimony in Irish High Court Case*. [Expert report]. Retrieved 1 October 2017, from <https://www.alston.com/en/resources/peter-swire-irish-high-court-case-testimony>
- Swire, P., Morell, M., Stone, G., Sunstein, C., & Clarke, R. (2013). *Liberty and security in a changing world. Report and Recommendations of The President's Review Group on Intelligence and Communications Technologies*. . Washington, DC: Executive Office of the President.
- Takes, F. (2014). *Algorithms for Analyzing and Mining Real-World Graphs* (Unpublished doctoral dissertation). Ph. D. dissertation, Leiden Institute of Advanced Computer Science (LIACS), Faculty of Science, Leiden University.
- Takes, F. (2018, September 9). *Re: Nieuwe versie (na bijeenkomst 6/9/2018)*. [Personal correspondence].
- Tawakol, O. (2016). *Letter to Customers and Partners*. [Blog-post]. Retrieved 9 April 2018, from <https://www.oracle.com/corporate/acquisitions/addthis/letter.html>
- The Tor Project (2015). *Tor web browser*. Cambridge, MA: The Tor Project, Inc. Retrieved 15 August 2015, from <https://www.torproject.org/>
- Theus, M., & Urbanek, S. (2008). *Interactive Graphics for Data Analysis: Principles and Examples*. Taylor & Francis.
- Thibault, C. (2013). *Game Data Analysis—Tools and Methods*. Birmingham: Packt Publishing Ltd.
- Thieurmél, B. (2016). *visNetwork: an R package for network visualization, using the vis.js javascript library* [Computer software manual]. Retrieved 1 march 2018, from <https://datastorm-open.github.io/visNetwork/>

- Thode, W., Griesbaum, J., & Mandl, T. (2015). 'I would have never allowed it': User Perception of Third-party Tracking and Implications for Display Advertising. In F. Pehar, C. Schlögl, & C. Wolff (Eds.), *Re: inventing Information Science in the Networked Society*. (pp. 445–456). Glückstad: Verlag Werner Hülsbusch.
- Thomas, C., Kline, J., & Barford, P. (2016). IntegraTag: a framework for high-fidelity web client measurement. In *Teletraffic Congress (ITC 28), 2016 28th international* (Vol. 1, pp. 278–285).
- Thomas, K. (1962). The structure of scientific revolutions.
- Thompson, K. (1968). Programming Techniques: Regular expression search algorithm. *Communications of the ACM*, 11(6), 419–422. doi: <http://dx.doi.org/0.1145/363347.363387>
- Tillmann, H. (2013). *Browser Fingerprinting: Tracking ohne Spuren zu hinterlassen*. Master's thesis. Berlin: Humboldt University, Institut für Informatike.
- Towvim, A. (2018, February 16). *Re: Re: DSP's and SSP's using tag management to track clicks / conversions?* [Personal correspondence].
- Traverso, S., Trevisan, M., Giannantoni, L., Mellia, M., & Metwalley, H. (2017). Benchmark and comparison of tracker-blockers: Should you trust them? In *Network Traffic Measurement and Analysis Conference (TMA), 2017* (pp. 1–9).
- Trevisan, M., Traverso, S., Metwalley, H., & Mellia, M. (2017). *Uncovering the Flop of the EU Cookie Law*. arXiv preprint arXiv:1705.08884.
- Tschofenig, H. (2013, July 11). *Writing Protocol Models*. [public-privacy (W3C mailing list)].
- Tschofenig, H. (2015, June 14). *Re: question on HTTP header fields*. [Personal correspondence].
- Tschofenig, H., & Van Eijk, R. (2011). Do Not Track, An Attempt to Frame the Debate. In *W3C workshop Web Tracking and User Privacy*. Princeton, NJ: W3C. [Position paper]. Retrieved 11 January 2014, from <http://www.w3.org/2011/track-privacy/papers/Tschofenig.pdf>
- Tukey, J. (1977). *Exploratory data analysis* (Vol. 2). Reading, Mass.
- Turcios Rodríguez, E. (2018). *Tracking cookies in the European Union, an empirical analysis of the current situation*. Master's thesis. Delft: Delft University of Technology.
- Unspam Technologies (n.d.). Project Honey Pot [Computer software manual]. Park City, Utah: Author. Retrieved 10 July 2015, from <http://www.projecthoneypot.org/faq.php>

- Van den Herik, H. J., & Van Eijk, R. (2013). Social Innovation & Big Data. In *IT Innovation Day*. Amersfoort: IDG Nederland. [Presentation].
- Van den Herik, H. J., & Van Eijk, R. (2014). The all-pervasive impact of Big Data in daily life: from Causality to Correlation. In *Launch of the Future Center - Smart Systems*. Leiden: Leiden Institute of Advanced Computer Science, Leiden University. [Presentation].
- Van den Herik, H. J., & Van Eijk, R. (2015). Big Data: een Introductie. In *Ondernemen en Big Data*. The Hague: VNO-NCW WEST. [Presentation].
- Van den Herik, H. J., & Van Eijk, R. (2016). De haakjes van Big Data. In *(Big) Data festival*. The Hague: Ministry of Infrastructure and Environment. [Presentation].
- Van der Hout, A. (2015, December 18). *Ad networks in the Netherlands*. [Personal correspondence].
- Van Eijk, R. (2011a). *Web Tracking Detection System (TDS)*. [Github]. Retrieved 1 January 2015, from <https://github.com/rvaneijk/TDS>
- Van Eijk, R. (2011b). *Web Tracking Detection System (TDS): An effective strategy to reduce systematic monitoring and profiling of user habits across websites*. MSc thesis. Leiden: Leiden University, Leiden Institute of Advanced Computer Science. Retrieved 5 March 2018, from <http://liacs.leidenuniv.nl/assets/Masterscripties/ICTiB/Rob-van-Eijk-non-confidential.pdf>
- Van Eijk, R. (2012). *The DNA of OBA: unique identifiers*. Retrieved 21 June 2015, from <http://web.archive.org/web/20141219090611/http://www.campusden Haag.nl/crk/publicaties/robvaneijk.html>
- Van Eijk, R. (2016a). *A brief introduction to Real Time Bidding (RTB)*. [Presentation]. Brussels: Article 29 Working Party. doi: <http://dx.doi.org/10.2139/ssrn.2856941>
- Van Eijk, R. (2016b). Informatie: Leidt meer informatie tot minder vrijheid? In A. in 't Groen, C. De Roon, & P. Slaman (Eds.), *Leids-Haags Allegorieënlexicon* (pp. 151–160). Delft: Uitgeverij Eburon.
- Van Eijk, R. (2017). Tracking classification: a cluster analysis approach to identify missing values. In N. Bielova, N. Laoutaris, A. Narayanan, & N. Nikiforakis (Eds.), (Vol. 7, p. 103). Dagstuhl, Germany: Schloss Dagstuhl–Leibniz-Zentrum für In-

- formatik. [Peer Reviewed Report]. doi: <http://dx.doi.org/10.4230/DagRep.7.4.83>
- Van Eijk, R. (2018). *Technology for Privacy: Protecting Against Online Tracking and Profiling*. [Presentation]. Brussels: Computers, Privacy, and Data Protection congress. doi: <http://dx.doi.org/10.2139/ssrn.3106582>
- Van Eijk, R., & Chester, J. (2014, July 28). *Recent Mergers & Acquisitions Involving Data Driven Marketing Companies*. [Personal correspondence].
- Van Eijk, R., & Chester, J. (2015, February 4). *RE: key mergers Jan 2013 - June 2014*. [Personal correspondence].
- Van Eijk, R., & Cortesi, A. (2016, August 24). *Tracking detection*. [Personal correspondence].
- Van Eijk, R., & Terra, F. (2012). *Cookie ID crawler*. [Github, private repository]. Retrieved 27 July 2015, from <https://github.com/rvaneijk/cookie-id-crawler>
- Van Eijk, R., & Terra, F. (2013). *Python script to read high entropy cookies from MITMDUMP and store these in a Neo4j property graph*. [Github, private repository]. Retrieved 16 August 2016, from <https://github.com/rvaneijk/cookie-id-crawler>
- Van Eijk, R., & Toubiana, V. (2016). *Python script to read high entropy cookies from OpenWPM and store these in a Neo4j property graph*. [Github]. Retrieved 16 August 2016, from <https://github.com/rvaneijk/openWPM-to-Neo4j>
- Van Kesteren, A., Gregor, A., Ms2ger, Russel, A., & Berjon, R. (2015). *W3C DOM4*. W3C Last Call Working Draft 18 June 2015. Retrieved 22 June 2015, from <http://www.w3.org/TR/dom/> (DOM 4)
- Van Goethem, T., & Joosen, W. (2017). One side-channel to bring them all and in the darkness bind them: Associating isolated browsing sessions.
- Vastel, A., Rudametkin, W., & Rouvoy, R. (2018). FP-TESTER: Automated Testing of Browser Fingerprint Resilience. In *IWPE 2018-4th International Workshop on Privacy Engineering* (pp. 1–5).
- Verhulst, D. (2012). *VMC-studiemiddag*. In *Mediaforum: Tijdschrift voor media- en communicatierecht*. Amsterdam: Cramwinckel.
- Vines, P., Roesner, F., & Kohno, T. (2017). Exploring ADINT: Using Ad Targeting for Surveillance on a Budget-or-How Alice Can Buy Ads to Track Bob. In *Proceedings of the 2017 on Workshop on Privacy in the Electronic Society* (pp. 153–164).

- Wagner, S., & Neshat, N. (2010). Assessing the vulnerability of supply chains using graph theory. *International Journal of Production Economics*, 126(1), 121–129.
- Wainberg, D. (2015, December 19). *Transitivity, DNT tracking status resource*. [Personal correspondence].
- Waldron, R., Pozdnyakov, M., & Shalamov, A. (2018). *Generic Sensor API*. W3C Candidate Recommendation, 20 March 2018. Retrieved 30 March 2018, from <https://www.w3.org/TR/2018/CR-generic-sensor-20180320/>
- Wang, J., Yuan, S., & Cai, K. (2015). *Real-Time Bidding based Display Advertising: Mechanisms and Algorithms*. *Real-Time Bidding A New Frontier of Computational Advertising Research*. [Presentation]. Padua: ACM International Conference on Web Search and Data Mining (WSD 2015). Retrieved 6 August 2016, from <http://tutorial-wsdm-2015.computational-advertising.org/images/WSDM2015.pdf>
- Wang, J., Yuan, S., & Zhang, W. (2016). *Real-Time Bidding based Display Advertising: Mechanisms and Algorithms*. [Presentation]. Padua: European Conference on Information Retrieval (ECIR 2016). Retrieved 6 August 2016, from <http://tutorial.computational-advertising.org/final-slides.pdf>
- Warren, S., & Brandeis, L. (1890). The right to privacy. *Harvard law review*, 193–220.
- Web Tracking and User Privacy. (2011, April 28-29). [Workshop report]. Princeton, NJ: W3C. Retrieved 11 April 2014, from <http://www.w3.org/2011/track-privacy/report.html>
- Wefers Bettink, W., Van Eijk, R., & Wagner, F. (2012). *Strictly Speaking: Cookies, Consent and Compliance*. Europe Data Protection congress. [Presentation]. Brussels: IAPP.
- Westin, A. (1967). *Privacy and freedom*. New York, NY: Atheneum.
- WHATWG (2015, December). *HTML Living Standard*. Last updated 11 December 2015. Retrieved from <http://html.spec.whatwg.org>
- Wheaton Lexomics (2012). *How to read a dendrogram*. [Video]. Wheaton College Massachusetts. Retrieved 18 February 2019, from <https://www.youtube.com/watch?v=MX6AUX1b1w0>
- White, N. (2014). *RNeo4j: interact with a Neo4j database in R*. [Github]. Retrieved 16 June 2015, from <http://nicolewhite.github.io/RNeo4j/docs/>
- White, T. (2012). *Hadoop: The definitive guide*. Sebastopol, CA: O'Reilly Media, Inc.

- Wilander, J. (2017). *Intelligent Tracking Prevention*. [Blogpost]. Retrieved 18 October 2017, from <https://webkit.org/blog/7675/intelligent-tracking-prevention/>
- Wilde, T., & Hess, T. (2007). Forschungsmethoden der wirtschaftsinformatik. *Wirtschaftsinformatik*, 49(4), 280–287.
- Wiley, S. (2017, May 2). *Re: congratulations and data set question*. [Personal correspondence].
- Wills, C. (2012, June 20). *WPM Privacy Measurement Bibliography follow-up*. [Personal correspondence].
- Yahoo (2002). *Yahoo! Search Marketing Advertiser Bidding Data, version 1.0*. [Dataset]. Retrieved 30 April 2017, from <https://webscope.sandbox.yahoo.com/catalog.php?datatype=a>
- Yahoo (2014). *Yahoo! Data Targeting User Modeling, Version 1.0*. [Dataset]. Retrieved 30 April 2017, from <https://webscope.sandbox.yahoo.com/catalog.php?datatype=a>
- Yonan, J. (2001). *OpenVPN*. Retrieved 2 December 2017, from <https://openvpn.net/index.php/open-source.html>
- Youyou, W., Kosinski, M., & Stillwell, D. (2015). Computer-based personality judgments are more accurate than those made by humans. *Proceedings of the National Academy of Sciences*, 112(4), 1036–1040.
- Yu, Z., Macbeth, S., Modi, K., & Pujol, J. M. (2016). Tracking the trackers. In *Proceedings of the 25th International Conference on World Wide Web* (pp. 121–132).
- Yuan, H., Maple, C., Chen, C., & Watson, T. (2018). Cross-device tracking through identification of user typing behaviours. *Electronics Letters*, 54(15), 957–959.
- Zevenbergen, B., Brown, I., Wright, J., & Erdős, D. (2013). Ethical Privacy Guidelines for Mobile Connectivity Measurements. *Available at SSRN 2356824*.
- Zhang, W., Yuan, S., & Wang, J. (2014). Optimal Real-Time Bidding for Display Advertising. In *Proceedings of the 20th ACM SIGKDD international conference on knowledge discovery and data mining*.
- Zhang, W., Yuan, S., Wang, J., & Shen, X. (2014). Real-Time Bidding Benchmarking with iPinYou Dataset. *arXiv preprint arXiv:1407.7073*.
- Zhou, Z., Diao, W., Liu, X., & Zhang, K. (2014). Acoustic fingerprinting revisited: Generate stable device id stealthily with inaudible sound. In *Proceedings of the 2014 ACM SIGSAC Conference on Computer and Communications Security* (pp. 429–440).

- Zorbas, K. (2015, December 12). *Ad networks in the Netherlands*. [Personal correspondence].
- Zuiderveen Borgesius, F. (2014). *Improving Privacy Protection in the Area of Behavioural Targeting*. PhD thesis. Amsterdam: University of Amsterdam.
- Zuiderveen Borgesius, F. (2015). Personal data processing for behavioural targeting: which legal basis? *International Data Privacy Law*, 5(3), 163.
- Zuiderveen Borgesius, F., & McDonald, A. M. (2015). Do Not Track for Europe. In *43rd Research Conference on Communication, Information and Internet Policy (Telecommunications Policy Research Conference)*. Available at SSRN: 2588086.

Appendices

To provide the reader with relevant background information, we list the appendices. They relate to various supplements to discussed subjects. The outline is as follows.

- A: Workshop participation
- B: Filtering challenging problems
- C: Acquisitions, mergers, and partnerships
- D: Example of partner networks: Denmark (DK)

A: Workshop participation

In Section 1.5 we marked the first step of our research methodology as workshop participation. In this appendix, we provide an overview. In Section A1 we present the different roles in which our participation took place. Section A2 contains an overview of the workshops that we have visited. All workshops contribute to Web Privacy Measurement (WPM).

A1: PARTICIPATING ROLES

During the time taken to conduct the research, Rob van Eijk participated in ten relevant workshops in four different roles.

- (1) Participation on behalf of the Autoriteit Persoonsgegevens.
- (2) Participation on behalf of the Article 29 Data Protection Working Party.
- (3) Participation on behalf of the Council of Europe.
- (4) Participation in a personal capacity.

Below, we give a brief characterization of the roles.

ROLE 1: AUTORITEIT PERSOONSGEGEVENS

Rob van Eijk served as Technologist and Senior Supervision Officer for the Autoriteit Persoonsgegevens, the Dutch DPA. In this role, he participated in four international working groups:

- (1) the Global Privacy Enforcement Network (GPEN),³⁶¹
- (2) the Internet Privacy Engineering Network (IPEN),³⁶²

361 GPEN was established „to foster cross-border co-operation among privacy authorities“. URL: <https://privacyenforcement.net/public/activities> (9 November 2015).

362 IPEN was established „to bring together developers and data protection experts with a technical background from different areas in order to launch and support projects that build privacy into everyday tools and develop new tools which can effectively protect and enhance our privacy“. URL: <https://secure.edps.europa.eu/EDPSWEB/edps/EDPS/IPEN> (9 November 2015).

- (3) the International Working Group on Data Protection in Telecommunication (IWGDPT),³⁶³ and
- (4) the Technology Subgroup of the Article 29 Data Protection Working Party (Art. 29 WP).³⁶⁴

ROLE 2: ARTICLE 29 DATA PROTECTION WORKING PARTY

Rob van Eijk represented the Article 29 Data Protection Working Party (Art. 29 WP) in various international meetings. His participation has been noted by, e. g., Boogert [2011], W. Davis [2013], Doty and Mulligan [2013, p. 152], Pels [2013], Libbenga [2013; 2014], Swift [2014], Zuiderveen Borgesius [2014, p. 332, footnote 1063], and Zuiderveen Borgesius and McDonald [2015, footnote 82 and acknowledgments]. Below, we list three representations.

- (1) the multi-stakeholder negotiations held by the World Wide Web Consortium (W3C) on Do Not Track (DNT),
- (2) a public debate on DNT held by the Member of the European Parliament (MEP) Amelia Andersdotter and MEP Françoise Castex, and
- (3) the Roundtable on Online Behavioral Advertising (OBA) held by the European Commission.³⁶⁵ The roundtable touched, e. g., on DNT as a potential consent mechanism in the European data protection and privacy context.

³⁶³ The International Working Group on Data Protection in Telecommunications (IWGDPT) was established in the framework of the International Conference of Data Protection and Privacy Commissioners (ICDPPC) which was established to create „an environment in which privacy and data protection authorities around the world are able effectively to act to fulfill their mandates, both individually and in concert, through diffusion of knowledge and supportive connections“. URL: <http://datenschutz-berlin.de/content/europa-international/international-working-group-on-data-protection-in-telecommunications-iwgdpt> (9 November 2015).

³⁶⁴ The Technology Subgroup of Art. 29 WP has a focus on addressing new technological developments that trigger (new) legal questions in respect to the EU data protection and privacy framework. The Art. 29 WP was set up „under the Directive 95/46/EC of the European Parliament and of the Council of 24 October 1995 on the protection of individuals with regard to the processing of personal data and on the free movement of such data“. URL: http://ec.europa.eu/justice/data-protection/article-29/index_en.htm (9 November 2015).

³⁶⁵ The roundtable was held by DG CONNECT, Sector Digital Privacy. URL: <https://ec.europa.eu/digital-agenda/sites/digital-agenda/files/Presentation%20Online%20Behavioural%20Advertising%20%28OBA%29.pdf> (9 November 2015).

ROLE 3: COUNCIL OF EUROPE

He represented the Council of Europe in the following workshop co-organized by the Council of Europe and the Internet Society. Internet Governance Forum, 'Who is following me: tracking the trackers'.³⁶⁶

ROLE 4: PERSONAL CAPACITY

In his personal capacity, he served as a member of the program committee of the following four workshops.

- (1) Berkeley Center for Law & Technology workshop Web Privacy Measurement (2012).³⁶⁷
- (2) W3C workshop Do Not Track and beyond (2012).³⁶⁸
- (3) W3C workshop on privacy and user-centric controls (2014).³⁶⁹
- (4) W3C workshop on privacy and linked data (2018).³⁷⁰

To provide somewhat more insights into the multi-stakeholder negotiations held by the W3C on DNT, we elaborate below on role 2 (Art. 29 WP). The W3C has kept an archive containing all the DNT discussions of the Tracking Protection Working Group on the public mailing list.³⁷¹

Furthermore, the Art. 29 WP has archived its two responses to two public calls for review by the W3C. The Technology Subgroup of Art. 29 WP drafted the letters (i. e., role 1). First, the response to W3C's public consultation of the technical W3C document containing the DNT building blocks [Article 29 Working Party, 2014a].³⁷² Second, the response to W3C's public consultation on the normative framework of DNT. [Article 29 Working Party, 2015b].³⁷³

³⁶⁶ URL: <http://www.internetsociety.org/who-following-me-tracking-trackers-igf2012> (9 November 2015).

³⁶⁷ URL: <https://www.law.berkeley.edu/centers/bclt/past-events/2012-conferences-2/may-2012-web-privacy-measurement/> (9 November 2015).

³⁶⁸ URL: <http://www.w3.org/2012/dnt-ws/> (9 November 2015).

³⁶⁹ URL: <http://www.w3.org/2014/privacyws/pc.html> (9 November 2015).

³⁷⁰ URL: <https://www.w3.org/2018/vocabws/program-committee.html> (20 April 2018).

³⁷¹ URL: <http://lists.w3.org/Archives/Public/public-tracking/> (9 November 2015).

³⁷² URL: http://ec.europa.eu/justice/data-protection/article-29/documentation/other-document/files/2014/20140606_wp29_ts_standardisation_letter_to_w3c.pdf (9 November 2015)

³⁷³ URL: http://ec.europa.eu/justice/data-protection/article-29/documentation/other-document/files/2015/20151001_letter_of_the_art_29_wp_w3c_compliance.pdf (9 November 2015).

A2: WORKSHOP OVERVIEW

This section provides an overview of the actual workshops. The overview groups the workshops together by each of the four roles (Section 12). Within a role, the workshops are grouped by hosting organization and sorted in alphabetical order. The year in which a workshop took place is shown in parentheses.

ROLE 1: AUTORITEIT PERSOONSgegevens

- (1) Conference Computers, Privacy & Data Protection (European Data Protection Supervisor (EDPS) panel 2015,³⁷⁴ Norwegian DPA panel 2016).³⁷⁵
- (2) European Commission (DG CONNECT): workshop SMART 2013/0071, ePrivacy Directive: assessment of transposition, effectiveness and compatibility with proposed Data Protection Regulation (2014).
- (3) Leiden University: e-Law symposium on the future of privacy and innovation (2015).
- (4) Future of Privacy Forum: New technologies - new privacy approaches? (2015).
- (5) Internet Privacy Engineering Network (2014).³⁷⁶
- (6) International Data Protection and Privacy Commissioners Conference (2015).
- (7) International Working Group on Data Protection in Telecommunications (2011 - 2016).
- (8) Institute for Information Law (IViR) & Berkeley Law workshop on online tracking protection & browsers (2011).
- (9) Institute for Information Law (IViR) & Berkeley Law workshop on online tracking (2014).
- (10) Leiden University and University of Twente: roundtable on privacy and WiFi tracking (2017).
- (11) Vereniging voor Media- en Communicatierecht Series: Aspects of cookies: web tracking and online behavioral advertising (2011).³⁷⁷

³⁷⁴ URL: <https://www.youtube.com/watch?v=6SEdnWLSZyk> (10 December 2015).

³⁷⁵ URL: <https://www.youtube.com/watch?v=BMcx07PRsKQ> (10 May 2016).

³⁷⁶ URL: <https://www.youtube.com/watch?v=m07bjmUAq-Q> (10 December 2015).

³⁷⁷ See also, Verhulst [2012].

ROLE 2: ARTICLE 29 DATA PROTECTION WORKING PARTY

- (1) European Parliament (LIBE): Browser settings and consent: A brief introduction to Do Not Track (DNT) (2017).
- (2) European Parliament: Do Not Track, is self-regulation enough? (2014).³⁷⁸
- (3) European Commission (DG CONNECT): Roundtable on online behavioral advertising (2013, 2014).
- (4) W3C Global Considerations Task Force Face-to-Face (2013).
- (5) W3C Tracking Protection Working Group, 2nd - 8th Face-to-Face (2011–2016).

ROLE 3: COUNCIL OF EUROPE

Internet Governance Forum: Workshop who is following me, tracking the trackers (2012).³⁷⁹

ROLE 4: PERSONAL CAPACITY

- (1) Amsterdam University: Amsterdam Privacy Conference (2012, 2018).
- (2) Amsterdam University: IViR Information Influx Conference (2014).
- (3) Berkeley Center for Law & Technology: Web Privacy Measurement Workshop (2012).
- (4) Conference Computers, Privacy & Data Protection (ENISA panel, 2018).³⁸⁰
- (5) Council of the European Union: Working Party on Telecommunications and Information Society (2017).
- (6) Datenschutzfachtagung workshop: Datenschutzmanagement à la Brüssel (2012).
- (7) Dagstuhl Seminar 17162 on Online Privacy and Web Transparency (2017).
- (8) Erasmus University: Leadership challenges with big data; workshop on ethical, legal and privacy challenges (2015, 2018).
- (9) Free and Open source Software Developers' European Meeting (2012, 2013, 2014).

³⁷⁸ URL: <https://www.youtube.com/watch?v=T9fqpE5e5S0> (10 December 2015).

³⁷⁹ URL: https://www.youtube.com/watch?v=mtPqt906u_g (10 December 2015).

³⁸⁰ URL: https://www.youtube.com/watch?v=_ehloemvJ0 (7 February 2018).

- (10) Future Of Privacy Forum on Personal Information: The benefits and risks of de-identification (2011).
- (11) IAPP Europe Data Protection Congress (2012) .
- (12) IAPP Practical Privacy Series: FTC & consumer privacy (2011).
- (13) League of European Research Universities: LERU Summer School (2016).
- (14) Leiden University: Big data symposium (Dual PhD Centre, Chamber of Commerce The Hague, VNO-NCW, 2015).
- (15) Leiden University: Big data symposium (Ministry of Infrastructure and Environment) (2016).
- (16) Leiden University: e-Law symposium on the future of privacy and innovation (2015).
- (17) Leiden University: Mini conference on the occasion of the opening of the Leiden Centre of Data Science (2014).
- (18) Oxford Internet Institute: workshop to discuss the new ethical, legal and technical challenges for research with vast datasets of mobile connectivity data (2013).³⁸¹
- (19) Princeton University: workshop on web privacy and transparency conference (livestream) (2014).
- (20) RSM, Erasmus University: Leadership challenges with big data & analytics (2015, 2016, 2018).
- (21) Wall Street Journal: Data transparency weekend (2012).³⁸²
- (22) Worchester Polytechnic Institute, Colloquium Series: Forensic aspects of digital investigation (2013).
- (23) W3C Workshop on Do Not Track and beyond (2012).
- (24) W3C Workshop on privacy and user-centric controls (2014).
- (25) W3C Workshop on web tracking and user privacy (2011).

³⁸¹ Acknowledgment in Zevenbergen, Brown, Wright, and Erdős [2013].

³⁸² Adams, Fedor, Felten, Felten, Fenton, Mayer, Ralli, Toubiana, Turner, and Van Eijk [2012] built a browser extension and named it Tracking Report Card. The browser extension showed end-users an indication of web tracking in a visual and easy-to-read reporting format. The report was collated from an automated database built with a cloud-based crawling infrastructure. Moreover, the data disclosed whether web sites honor opt-out and DNT requests by removing its tracking cookies. The browser extension provided non-technical end-users with information about whether and how they were tracked.

B: Filtering challenging problems

In Section 1.5 we marked the second step of our methodology as filter the challenging problems put forward in the workshops in a way that we can combine the problems resulting in distilling a PS and the corresponding RQs. Table B.1 shows the result of the second step of our research methodology. For completeness we remark that Table 1.1 links the PS and the RQs to the precise places where we address RQ1 and RQ2.

Table B.1: Mapping of the filtered workshops with RQs.

FILTERED WORKSHOP	RQ1	RQ2
Berkeley Law Web Privacy Measurement Workshop	✓	
EU Parliament workshop Do Not Track, is self-regulation enough?		✓
EU Roundtable Online Behavioral Advertising		✓
Dagstuhl Seminar 17162 Online Privacy and Web Transparency	✓	✓
IGF Workshop Tracking the Trackers		✓
IViR & Berkeley Law Workshop on Online Tracking		✓
Princeton Web Privacy and Transparency Conference	✓	
W3C Tracking Protection Working Group meetings	✓	✓
W3C Workshop Do Not Track and beyond		✓
W3C Workshop Privacy and User-Centric Controls		✓
W3C Workshop Web Tracking and User Privacy		✓

C: Acquisitions, mergers, and partnerships

Table C.1: Acquisitions and mergers (1/7) [Van Eijk & Chester, 2014; 2015].

TECHNOLOGY CONSOLIDATION	(JANUARY 2013 - DECEMBER 2015)
Accenture [Acquity Group]	AOL [PrecisionDemand]
Accenture [AD.Dialetto]	AOL [velos]
Accenture [Chaotic Moon Studios]	AOL [Vidible]
Accenture [Cimation]	AppNexus [Alenty]
Accenture [Cloud Sherpas]	AppNexus [MediaGlu]
Accenture [i4C Analytics]	AppNexus [Yieldex]
Accenture [Mortgage Cadence]	Berkshire Partners [Catalina]
Accenture [Procurian]	Brightcove [Unicorn Media]
Accenture [Reactive Media]	Clarke Holdings [Valassis]
Accenture [Sagacious Consultants]	Comcast [Contingent Netw. Serv.]
Accenture [Tquila]	Comcast [FreeWheel]
Axiom Corp [Allant Group]	Comcast [NBC Universal]
Axiom Corp [LiveRamp]	Comcast [PowerCloud Systems]
Adobe [Aviary]	Comcast [Visible World]
Adobe [Digital Analytix]	comScore [MdotLabs]
Adobe [Efficient Frontier]	comScore [Proximic]
Adobe [Fotolia]	comScore [Rentrak Corporation]
Adobe [Ideacodes]	Criteo [AdQuantic]
Adobe [Mixamo]	Criteo [DataPop]
Adobe [Neolane]	Criteo [Tedemis]
Adobe [Satellite]	DataXu [JasperLabs]
Adobe [Thumb Labs]	Deloitte Digital [Banyan Branch]
AOL [Adap.tv]	Dentsu [Explido]
AOL [Aiguentacar]	Dentsu [Fifty Four Media]
AOL [Convertro]	Dentsu [IID]
AOL [gdgt]	Dentsu [Lesmobilizers]
AOL [Gravity]	Dentsu [Oddfellows]
AOL [Kanvas Labs]	Dropbox [Aria Glassworks]
AOL [Millennial Media]	Dropbox [Bubbli]

Table C.2: Acquisitions and mergers - continued (2/7).

TECHNOLOGY CONSOLIDATION	(JANUARY 2013 - DECEMBER 2015)
Dropbox [Clementine]	Facebook [Monoidics]
Dropbox [CloudOn]	Facebook [Oculus VR]
Dropbox [Droptalk]	Facebook [Onavo]
Dropbox [Endorse]	Facebook [Osmeta]
Dropbox [Foundry Hiring]	Facebook [Parse]
Dropbox [HackPad]	Facebook [PrivateCore]
Dropbox [Loom]	Facebook [ProtoGeo]
Dropbox [Mailbox]	Facebook [Pryte]
Dropbox [MobileSpan]	Facebook [QuickFire Networks]
Dropbox [Parastructure]	Facebook [Spaceport.io]
Dropbox [Photo Lab]	Facebook [SportStream]
Dropbox [PiCloud]	Facebook [Storylane]
Dropbox [Pixelapse]	Facebook [Teehan & Lax]
Dropbox [Predictive Edge]	Facebook [TheFind, Inc.]
Dropbox [Readmill]	Facebook [VocalIQ]
Dropbox [Sold]	Facebook [WhatsApp]
Dropbox [TapEngage]	Facebook [Wit.ai]
Dropbox [Umano]	Google [Adometry]
Dropbox [Zulip]	Google [Agawi Inc]
Dunnhumby [Sociomantic Labs]	Google [Alpental Technologies]
Dunnhumby [Standard Analytics]	Google [Appetas]
eXelate [Affinova]	Google [Appurify]
eXelate [Arbitron]	Google [Autofuss]
eXelate [eXelate]	Google [Bebop]
eXelate [Harris Interactive]	Google [Behavio]
eXelate [Indicus Analytics]	Google [Bitspin]
Experian [41 st Parameter]	Google [Boston Dynamics]
Experian [AdTruth]	Google [Bot & Dolly]
Experian [Conversen]	Google [Bump Technologies]
Facebook [Ascenta]	Google [Dark Blue Labs]
Facebook [Atlas]	Google [Deepmind Technolgies]
Facebook [Branch Media]	Google [Digisfera]
Facebook [Endaga]	Google [Directr]
Facebook [Hot Studio]	Google [Divide]
Facebook [Jibbig]	Google [Divshot]
Facebook [Little Eye Labs]	Google [DNNresearch]
Facebook [LiveRail]	Google [drawElements]

Table C.3: Acquisitions and mergers - continued (3/7).

TECHNOLOGY CONSOLIDATION	(JANUARY 2013 - DECEMBER 2015)
Google [Emu Messenger]	Google [Timeful]
Google [Firebase]	Google [Titan Aerospace]
Google [FlexyCore]	Google [Toro]
Google [Flutter]	Google [Vision Factory]
Google [Fly Labs]	Google [Wavii]
Google [Gecko Design Inc.]	Google [Waze]
Google [Green Throttle Games]	Google [WIMM Labs]
Google [Holomni]	Google [Zync Render]
Google [Imperium]	Gravity4 [adX Search]
Google [Industrial Perception Inc.]	Gravity4 [Argyle Social]
Google [Jetpac]	Gravity4 [Bolzter]
Google [Jibe Mobile]	Gravity4 [Conyak]
Google [Launchpad Toys]	Gravity4 [EuroAds]
Google [Lift Labs]	Gravity4 [EzLike]
Google [mDialog]	Gravity4 [Kanary]
Google [Meka Robotics]	Gravity4 [Pixels]
Google [Nest Labs]	Gravity4 [Triggit]
Google [Odyssey]	Gravity4 [Zurmo]
Google [Oyster]	Group M [Plista]
Google [Pie]	Group M [The Exchange Lab]
Google [Pixate]	Havas Media [ElisaDBI]
Google [Polar]	Havas Media [GT Media]
Google [Pulse.io]	Havas Media [MFG Labs]
Google [Quest Visual]	Havas Media [Revenue Frontier]
Google [Rangespan]	Hitachi Data Syst. [Pentaho]
Google [RelativeWave]	Hitachi Data Syst. [Prizm Payment Serv.]
Google [Schaff]	IBM [AlchemyAPI]
Google [Skillman & Hackett]	IBM [Aspera]
Google [Skybox Imaging]	IBM [Blekko]
Google [SlickLogin]	IBM [Blue Box]
Google [Softcard]	IBM [Clearleap]
Google [Songza]	IBM [Cleversafe]
Google [Spider.io]	IBM [Cloudant]
Google [Stackdriver]	IBM [Cognea]
Google [Talaria]	IBM [Compose]
Google [Terra Bella]	IBM [CrossIdeas]
Google [Thrive Audio]	IBM [CSL International]

Table C.4: Acquisitions and mergers - continued (4/7).

TECHNOLOGY CONSOLIDATION	(JANUARY 2013 - DECEMBER 2015)
IBM [Daeja Image Systems]	Lotame [AdMobius]
IBM [Explorys]	Mastercard [Applied Predictive Techn.]
IBM [Fiberlink]	Mastercard [C-sam]
IBM [Gravitant]	Mastercard [Pinpoint]
IBM [Merge Healthcare]	Mastercard [Provus]
IBM [Meteorix]	Mastercard [Transaction Network Serv.]
IBM [Phytel]	MediaMath [Akamai]
IBM [Silverpop]	MediaMath [Spree7]
IBM [SoftLayer]	MediaMath [Tactads]
IBM [Star Analytics]	Merkle [50ofriends]
IBM [StrongLoop]	Merkle [5th Finger]
IBM [The Now Factory]	Merkle [Brilig]
IBM [The Weather Company]	Merkle [IMPAQT]
IBM [Trusteer]	Merkle [New Control]
IBM [Unica]	Merkle [RKG]
IBM [Urbancode]	Milennial [Jumtap]
IBM [Xtify]	Milennial [Metaresolver]
IgnitionOne [Knotice]	Milennial [NEXAGE]
Interpublic [Black Sheep]	Neustar [Aggregate Knowledge]
Interpublic [Genuine Interactive]	Neustar [Bombora Technologies]
Interpublic [Halesway]	Neustar [GO.CO]
Interpublic [Perfect Fools]	Neustar [MarketShare]
Interpublic [Profero]	NICE Systems [Causata]
Interpublic [Spookes Labs]	Omnicom [Agency Republic]
Kroger [Dunnhumby]	Omnicom [Haygarth]
Kroger [Harris Teeter]	Omnicom [HMKM]
Kroger [Roundy's]	Omnicom [LiThe]
Kroger [Vitacost]	Omnicom [Media Interactive]
LinkedIn [Bizo]	Omnicom [Mobile5 Media]
LinkedIn [Bright.com]	Oracle [Acme Packet]
LinkedIn [Careerify]	Oracle [AdKarma]
LinkedIn [Fliptop]	Oracle [All Media Network]
LinkedIn [lynda.com]	Oracle [BigMachines]
LinkedIn [Newsle]	Oracle [Bitzer Mobile]
LinkedIn [Pulse]	Oracle [BlueKai]
LinkedIn [Refresh.io]	Oracle [CloudMonkey]
Lithium [Klout]	Oracle [Compendium]

Table C.5: Acquisitions and mergers - continued (5/7).

TECHNOLOGY CONSOLIDATION	(JANUARY 2013 - DECEMBER 2015)
Oracle [Corente]	Salesforce [Kerensen Consulting]
Oracle [Datalogix]	Salesforce [MinHash]
Oracle [Eloqua]	Salesforce [Pardot]
Oracle [Front Porch Digital]	Salesforce [RelateIQ]
Oracle [GreenBytes]	Salesforce [SteelBrick]
Oracle [LiveLOOK]	Salesforce [Tempo AI]
Oracle [LYFE Mobile]	Salesforce [Toopher]
Oracle [Maxymiser]	SDL [Alterian]
Oracle [Micros Systems]	Sizmek [PointRoll]
Oracle [Nimbula]	Sizmek [StrikeAd]
Oracle [Push IO]	Spredfast [Mass Relevance]
Oracle [Responsys]	Spredfast [Shoutlet]
Oracle [Rhythm NewMedia]	Sprinklr [Booshaka]
Oracle [StackEngine]	Sprinklr [Branderati]
Oracle [Tekelec International]	Sprinklr [Dachis Group]
Oracle [TOA Technologies]	Sprinklr [Get Satisfaction]
Pitney Bowes [Borderfree]	Sprinklr [newBrandAnalytics]
Pitney Bowes [Real Time Content]	Sprinklr [Pluck]
PubMatic [Mocean Mobile]	Sprinklr [Scup]
Rakuten Marketing [DC Storm Limited]	Sprinklr [TBG Digital]
Rakuten Marketing [Deep Forest Media]	Staples [Accolade Promotion Group]
Rakuten Marketing [Ebates]	Staples [Office Depot]
Rakuten Marketing [Fits.me]	Staples [PNI Digital Media]
Rakuten Marketing [Overdrive]	Staples [Runa]
Rakuten Marketing [Slice]	Teradata [Appoxee]
Rakuten Marketing [Spotlight]	Teradata [FLXone]
Rakuten Marketing [Viber Media]	Teradata [Hadapt]
Rakuten Marketing [Viki]	Teradata [RainStor]
Rakuten Marketing [Voyagin]	Teradata [Revelytix]
Rakuten Marketing [Webgistix]	Teradata [Think Big Analytics]
RichRelevance [Precog]	Twitter [Adrenaline Mobility]
RocketFuel [X+1]	Twitter [Apps & Zerts]
Rubicon Project [Chango]	Twitter [Bluefin Labs]
Salesforce [Clipboard]	Twitter [CardSpring]
Salesforce [EdgeSpring]	Twitter [Cover Lockscreen]
Salesforce [EntropySoft]	Twitter [Crashlytics]
Salesforce [ExactTarget]	Twitter [Fastlane]

Table C.6: Acquisitions and mergers - continued (6/7).

TECHNOLOGY CONSOLIDATION (JANUARY 2013 - DECEMBER 2015)	
Twitter [Gnip]	Xaxis [24/7 Media]
Twitter [Lucky Sort]	Xaxis [Bannerconnect]
Twitter [Marakana]	Yahoo [Admovate]
Twitter [Mesagraph]	Yahoo [alike]
Twitter [Mitro]	Yahoo [Astrid]
Twitter [MoPub]	Yahoo [Aviate]
Twitter [Namo Media]	Yahoo [Bignoggins Productions]
Twitter [Niche]	Yahoo [Blink]
Twitter [Periscope Co]	Yahoo [Bookpad]
Twitter [SecondSync]	Yahoo [Bread]
Twitter [SnappyTV]	Yahoo [BrightRoll]
Twitter [Spindle]	Yahoo [ClarityRay]
Twitter [TapCommerce]	Yahoo [Cloud Party]
Twitter [Tellapart]	Yahoo [Cooliris]
Twitter [tenXer]	Yahoo [Distill]
Twitter [Trendrr]	Yahoo [EvtLive]
Twitter [Ubalo]	Yahoo [Flurry]
Twitter [Whetlab]	Yahoo [GhostBird Software]
Twitter [ZeroPush]	Yahoo [GoPollGo]
Twitter [Zipdial]	Yahoo [Hitpost]
Verizon [OnCue]	Yahoo [Incredible Labs]
Walmart Labs [Adchemy]	Yahoo [IQ Engines]
Walmart Labs [Inkiru]	Yahoo [Jybe]
Walmart Labs [Luvocracy]	Yahoo [Lexity]
Walmart Labs [PunchTab]	Yahoo [Loki Studios]
Walmart Labs [Stylr]	Yahoo [LookFlow]
Walmart Labs [Tasty Labs]	Yahoo [Luminate]
Walmart Labs [Torbit]	Yahoo [Media Group One]
WPP [DNX]	Yahoo [Meh Labs]
WPP [FusePump]	Yahoo [MessageMe]
WPP [Medialets]	Yahoo [MileWise]
WPP [Ootworld]	Yahoo [PeerCDN]
WPP [Percolate Industries]	Yahoo [PlayerScale]
WPP [Precise Media]	Yahoo [Polyvore]
WPP [The Data Republic]	Yahoo [Propeld]
WPP [Twist Image]	Yahoo [Ptch]
WPP [VML]	Yahoo [Quik.io]

Table C.7: Acquisitions and mergers - continued (7/7).

TECHNOLOGY CONSOLIDATION (JANUARY 2013 - DECEMBER 2015)	
Yahoo [Qwiki]	Yahoo [Tomfoolery]
Yahoo [RayV]	Yahoo [Tumblr]
Yahoo [Rockmelt]	Yahoo [Vizify]
Yahoo [Rondee]	Yahoo [Wander]
Yahoo [SkyPhrase]	Yahoo [Xobni]
Yahoo [Sparq]	Yahoo [Zofari]
Yahoo [SteamFunk Labs]	Yahoo [Ztelic]
Yahoo [Summly]	

D: Partner networks: Denmark (DK)

Here, we provide somewhat more complex examples of partner networks in Denmark (DK), which is situated on the right hand of the scale (see Figure 4.6).³⁸³ The aim of the examples is to zoom in on the five leading partner networks and their interconnections.

The results reported in this appendix are based on research data collected by visiting 475 news items from 25 news websites in Denmark. Table D.1 lists the number of news items (denoted as ‘items’ in the second and fourth column) per digital media (denoted as ‘origin’ in the first and third column).

Table D.1: Overview of news websites in Denmark (DK).

ORIGIN	ITEMS	ORIGIN	ITEMS
bornholm.nu	38	rudersdal.lokalavisen.dk	19
nordjyske.dk	38	stiften.dk	19
www.fyens.dk	38	tidende.dk	19
albertslund.lokalavisen.dk	19	www.bt.dk	19
cphpost.dk	19	www.jv.dk	19
ekstrabladet.dk	19	www.kristeligt-dagblad.dk	19
folketidende.dk	19	www.information.dk	19
furesoe.lokalavisen.dk	19	borsen.dk	16
gentofte.lokalavisen.dk	19	jyllands-posten.dk	14
hilleroed.lokalavisen.dk	19	www.jyllands-posten.dk	4
hoersholm.lokalavisen.dk	19	penge.borsen.dk	3
lyngby-taarbaek.lokalavisen.dk	19	finans.dk	1
politiken.dk	19		

I compiled and refactored a referrer graph from our research data collected from Denmark. The resulting graph is depicted in Figure D.1. The node ‘crawled.io’ is depicted as the center of the blue subgraph. We made the graph sparse with a directed-force

³⁸³ Denmark (DK) with $N=719$ nodes, $E=1,348$ edges, and $C_B=33$ clusters.

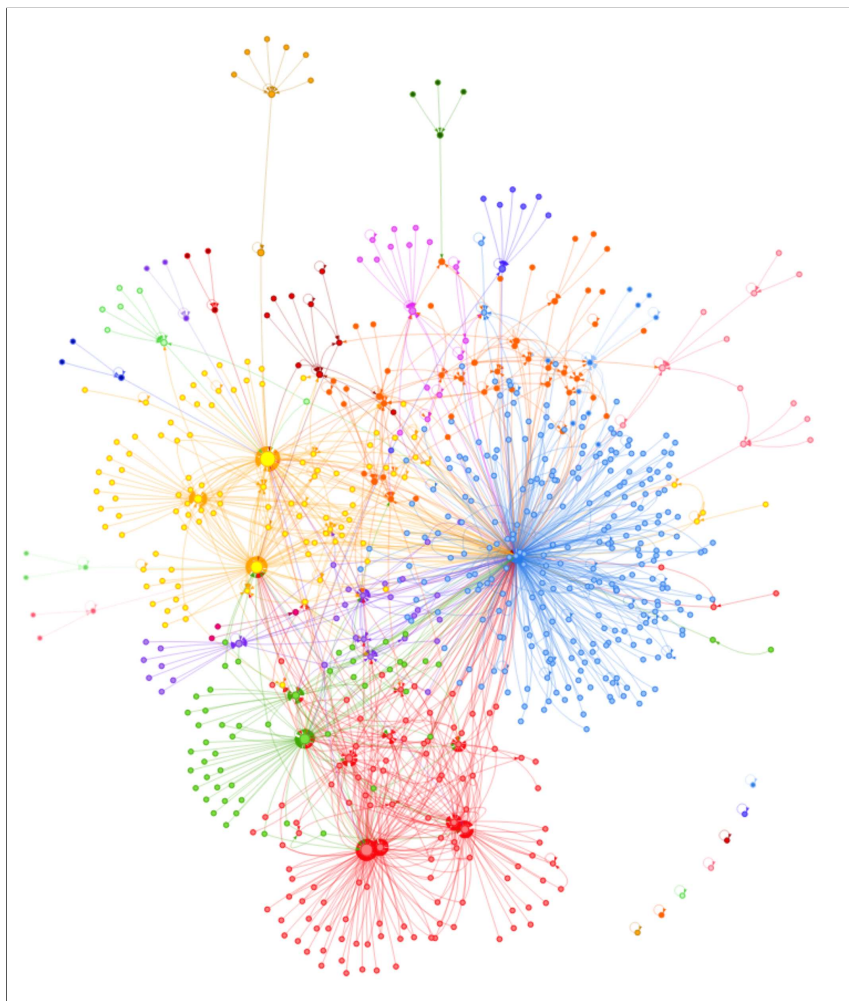


Figure D.1: Generic web-tracking graph: Denmark (DK) with $N=719$ nodes, $E=1,348$ edges, and $C_B=33$.

layout. The directed graph consists of 719 distinct nodes and 1,348 distinct edges. The following seven nodes are not connected to any other node:

- (1) 'o.twimg.com',³⁸⁴
- (2) 'pbs.twimg.com',
- (3) 'ton.twimg.com',
- (4) 'aarhusfestuge.dk',³⁸⁵
- (5) 'realtime.services.discuss.com',³⁸⁶
- (6) 'livewire.ritzau.de:6101',³⁸⁷ and
- (7) 'aus5.mozilla.org'.³⁸⁸

We grouped these unconnected nodes together in the bottom-right corner of Figure D.1.

We remark that 33 clusters of partner networks were identified with a cluster-edge betweenness algorithm [Csárdi & Nepusz, 2006]. Below we list the five leading clusters identified by their central node(s) (see Definition 4.16).

- (1) the Rubicon Project (Figure D.2, p. 269),
- (2) Adform (Figure D.3, p. 270),
- (3) Adspine (Figure D.4, p. 271),
- (4) AppNexus (Figure D.5, p. 272), and
- (5) Google (Figure D.6, p. 273).

Interconnection (see Definition 4.19) between partner network is depicted nicely in, e. g., Figure D.5. The picture shows precisely where AppNexus is connected with two edges into each of its neighboring networks, i. e., the Rubicon Project (see Figure D.2), Adform (see Figure D.3), and Google (see Figure D.6).

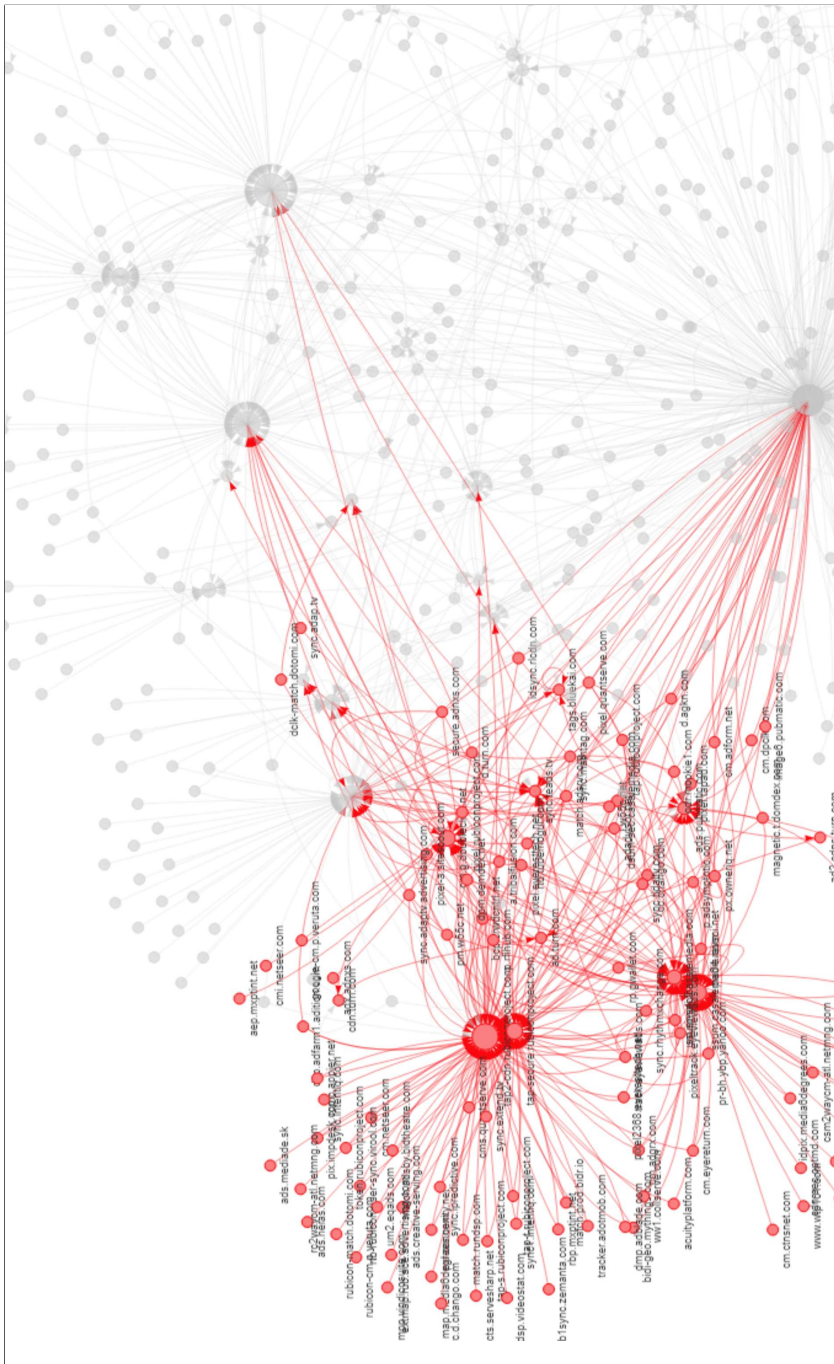
³⁸⁴ The origins 'twimg.com' belongs to Twitter.

³⁸⁵ This seems to be an anomaly, i. e., an HTTP request without a referrer header or a location header.

³⁸⁶ HTTP traffic related to the Disqus blog comment hosting service. URL: <https://www.crunchbase.com/organization/disqus> (24 February 2018).

³⁸⁷ HTTP traffic related to the Livewire social platform. URL: <https://www.crunchbase.com/organization/livewire> (24 February 2018).

³⁸⁸ The origin 'aus5.mozilla.org' is related to the headless crawler.



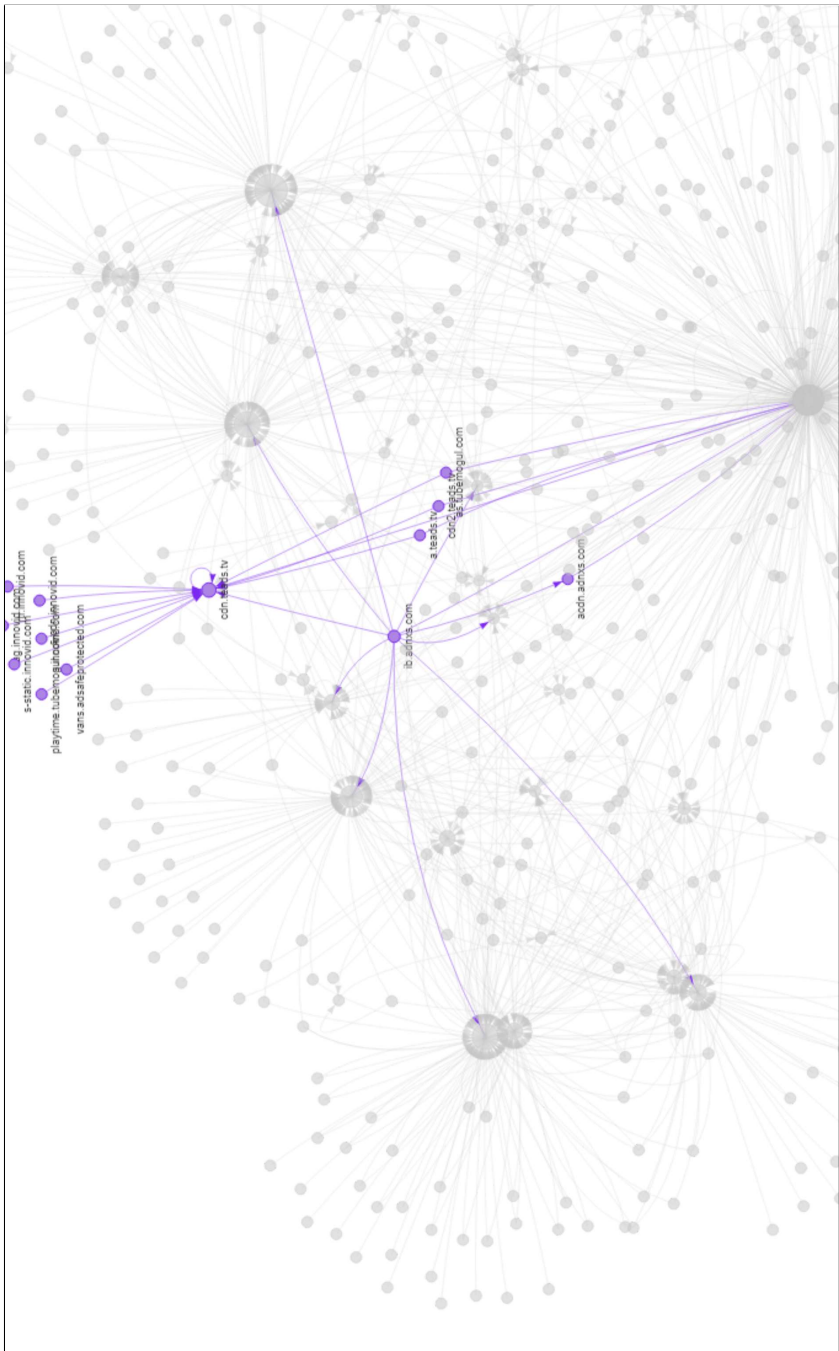


Figure D.5: Partner network: AppNexus (DK).

Summary

WEB PRIVACY MEASUREMENT IN REAL-TIME BIDDING SYSTEM A GRAPH-BASED APPROACH TO RTB SYSTEM CLASSIFICATION

Web Privacy Measurement (WPM) is the observation of websites and services to detect, characterize, and quantify privacy impacting behaviors [Englehardt et al., 2014]. Real-Time Bidding (RTB) is a big data application within the organizational field of marketing to improve sales by real-time data-driven marketing and personalized (behavioral) advertising. The practice of RTB manifests itself on, e. g., news websites.

The aim of the thesis is to investigate and understand to what extent we can measure the privacy component connected to RTB. In particular, we focus on two related research questions: (1) how do we move from data collection to graph analysis? and (2) what are the emerging characteristics of the graph which are fit for graph analysis? Below we provide first a broad overview of the thesis and then we give a brief chapter-by-chapter description of the contents.

For our investigations, we crawled articles from European news websites with the help of the Netograph experimental framework. To collect specific research data in this context we adopted the assumption that today online news continues to rely on advertising as a business model.

Next, we filtered metadata from the HTTP headers of the crawl data. Then we built a small-data model with metadata derived from the web crawls. The small-data model expresses the interrelationships between RTB actors. We stored the network of RTB partners in a separate graph database and did so by EU country. We then refactored the graph databases by grouping the origins (domain names) of all the news items that we crawled on one single node. Subsequently, we applied two types of network science algorithms to our databases, i. e., (1) cluster-edge betweenness and (2) node betweenness.

First, cluster-edge betweenness turned out to be a strong indicator for the identification of the main actor(s) in a partner network.

Second, node betweenness showed to be a strong indicator for networks of RTB partners. The application of the two algorithms enabled us to categorize the nodes (i. e., actor(s) in an RTB-partner network) into four types:

- (1) bridging node,
- (2) interconnecting node,
- (3) central node, and
- (4) satellite node.

The first two types are the result of the application of cluster-edge betweenness. The third and fourth type are the result of the application of node betweenness.

Based on the experiments we arrived at three conclusions. First, cluster-edge betweenness and node betweenness help us in understanding the interconnectedness of the ad-technology companies. Second, we may conclude that the *combined* application of both network science algorithms to empirical WPM-data enables us to categorize a partner network. Third, the research also shows that the interconnection in a betweenness cluster is caused by the data flows of the companies themselves due to their specializations in ad technology. This is the main reason why categorization of RTB-partner networks with cluster-edge betweenness and node betweenness algorithms is effective.

In Chapter 1, we provide a background for our investigation emphasizing the topics privacy and RTB systems. We start by briefly looking at data, big data, small data, and metadata. Then we provide our definition of web tracking and justify it by arguments. Our definition of web tracking is meant to position the thesis within the larger privacy debate. We present our problem statement and formulate two research questions. Furthermore, we delineate the research methodology to answer the research questions.

In Chapter 2, we review literature on Web Privacy Measurement (WPM). We do so by providing a description of (1) three client identification mechanisms, (2) deterministic and probabilistic identification, and (3) web-based fingerprinting. The descriptions serve as a background for the state of the art in web tracking. Furthermore, we briefly describe HTTP header fields. We close the chapter by providing an overview of (1) four different approaches to data collection in the academic field of WPM and (2) nine WPM visualization tools.

In Chapter 3, we develop our novel Graph-Based Methodological Approach (GBMA) to identify web tracking in HTTP header fields (flow data). Our approach consists of three parts: data collection (Part 1), data reduction (Part 2), and data modeling (Part 3). We conduct experiments from an end-user perspective. Of course, we retain the research data. For this purpose, we describe how we reduced the big data in such a way that we can perform our experiments with metadata from the HTTP header. We show that *entropy* as a measure fails to differentiate between tracking cookies and non-tracking cookies. Therefore, we model the context of online media with specific small-data elements. Then we provide a metadata model (a small-data model) and store the data into a graph.

In Chapter 4, we present four views on RTB systems. First, we present an *empirical* view of RTB as a network of partners deployed on national and regional European news websites. Then, we give a *theoretical* view of the key concepts in RTB systems reflected in partner networks. Thereafter, we discuss a *legal* view. Here we identify differences in consent implementations for tracking cookies in EU countries. To do so, we rank the countries on a scale. From the results we arrive at two conclusions: (1) EU countries are dominated by strict-consent mechanisms on the left end of the scale and (2) EU countries where implied-consent mechanisms have the upper hand are to be found on the right end of the scale. These two conclusions are supported by applying the GBMA with the network science algorithms to our contextual research data. Finally we completed our research by a *societal* view. Our main finding is that the roles of the companies involved with web tracking can be explained through expressing metadata from the HTTP header into (mathematical) characteristics of nodes and edges of the graph.

In Chapter 5 we provide an answer to the problem statement and there we formulate the conclusions of the thesis. Then we recommend directions for further research.

In summary, this thesis makes the following three contributions. First, to the best of our knowledge, it is the first (empirical) study to apply a Graph-Based Methodological Approach (GBMA) to network science algorithms for measuring the privacy component of RTB. Second, we describe a general method of categorizing RTB systems based on cluster-edge betweenness and node-betweenness

algorithms. Third, we demonstrate our GBMA (approach) on a dataset of European national and regional news websites.

Samenvatting (Summary in Dutch)

WEB PRIVACY MEASUREMENT IN REAL-TIME BIDDING SYSTEMEN EEN GRAAFGEBASEERDE AANTAK VOOR DE CLASSIFICATIE VAN RTB SYSTEMEN

Web Privacy Measurement (WPM) is het observeren van websites en services met als doel om (mogelijke) inbreuken op de privacy (rechten en vrijheden) van betrokkenen te detecteren, karakteriseren en kwantificeren (cf. [Englehardt et al., 2014]). Real-Time Bidding (RTB) is een *big-data* toepassing binnen de marketingactiviteiten van een organisatie met als doel om de verkoop te verbeteren door middel van *real-time* datagedreven marketing en gepersonaliseerde (gedrags)advertenties. De praktijk van RTB manifesteert zich bijvoorbeeld op websites van Europese kranten en andere digitale media.

Het doel van mijn proefschrift is om te onderzoeken in hoeverre we de privacycomponent die verbonden is met RTB kunnen meten. We richten ons daarbij met name op twee onderzoeksvragen: (1) hoe gaan we van een dataverzameling naar graafanalyse? en (2) wat zijn de emergente karakteristieken van de graaf die geschikt zijn voor graafanalyses?

Voor ons onderzoek zijn nieuwsartikelen bezocht op Europese krantenwebsites op een geautomatiseerde wijze met behulp van de Netograph-webcrawldienst. Het volledige netwerkverkeer tussen de browser en de website is daarbij vastgelegd. De belangrijkste reden voor het verzamelen van onderzoeksgegevens in deze context is gebaseerd op de aanname dat het businessmodel van krantenwebsites nog steeds (sterk) afhankelijk is van inkomsten uit advertenties.

Vervolgens zijn HTTP-headers als metadata uit de ruwe data gefilterd. Hiermee kan een small-data model (dat is een gerichte subgraaf) gemaakt worden. Het small-data model drukt de onderlinge relaties uit tussen RTB-actoren, die in een afzonderlijke grafendatabase per EU-land zijn opgeslagen. De grafendatabase is geherstructureerd door het oorsprongpunten (domeinnamen) van alle digitale nieuwsitems die zijn bezocht te groeperen op

één enkel knooppunt. Daarbij zijn twee soorten network science algoritmen toegepast op onze database te weten (1) cluster-edge-betweenness en (2) node-betweenness.

In de context van RTB blijkt cluster-edge-betweenness een sterke indicator te zijn voor de identificatie van de belangrijkste actor(en) in een RTB-partnernetwerk. Node-betweenness blijkt een sterke indicator te zijn voor netwerken van RTB-partners. Door toepassing van de twee algoritmen kunnen we de knooppunten (d.w.z. de actor(en) in een RTB-partnernetwerk) categoriseren in vier typen:

- (1) overbruggingsknooppunt,
- (2) verbindend knooppunt,
- (3) centraal knooppunt en
- (4) satellietknooppunt.

De eerste twee typen zijn het resultaat van de toepassing van cluster-edge-betweenness. Het derde en vierde type zijn het resultaat van de toepassing van node-betweenness.

Uit de experimenten mogen we drie conclusies trekken. Ten eerste, cluster-edge-betweenness- en node-betweenness-algoritmen stellen ons in staat om de onderlinge verbondenheid van de advertentietechnologiebedrijven beter te begrijpen. Ten tweede, de gecombineerde toepassing van zowel network science algoritmen als empirische WPM-gegevens stellen ons in staat om een netwerk van RTB-partners te categoriseren. Ten derde toont het onderzoek ook aan dat de interconnectie in een betweenness-cluster veroorzaakt wordt door de datastromen van de bedrijven zelf vanwege hun specialisaties in advertentietechnologie. Dit is de belangrijkste reden waarom het categoriseren van netwerken van RTB-partners met cluster-edge-betweenness- en node-betweenness-algoritmen effectief is.

In hoofdstuk 1 wordt een achtergrond geschetst van het onderzoek in privacy- en RTB-systemen. Het verhaal begint met een kort overzicht van de begrippen gegevens: big data, small data en metadata. Tevens wordt een definitie gegeven voor de term webtracking (het gevolgd worden op internet). We motiveren deze concepten met argumenten. Vervolgens beoog ik met de definitie voor webtracking het proefschrift in een breder debat te plaatsen over privacy. Ik presenteer de probleemstelling en formuleer twee onderzoeksvragen. Verder wordt de onderzoeks-methodologie uiteen gezet.

In hoofdstuk 2 bespreken we de literatuur over Web Privacy Measurement (WPM). We geven een beschrijving van (1) drie client-identificatiemechanismen, (2) deterministische en probabilistische identificatie en (3) webgebaseerde vingerafdrukken. Deze beschrijvingen dienen als achtergrond voor de stand van zaken in webtracking, gevolgd door een korte beschrijving van een aantal velden in de HTTP-header. We sluiten het hoofdstuk af met een overzicht van (1) vier verschillende benaderingen van dataverzameling binnen het vakgebied van WPM en (2) negen visualisatiehulpmiddelen van WPM.

In hoofdstuk 3 staat de ontwikkeling van onze nieuwe GraafgeBaseerde Methodologische Aanpak GBMA centraal. Het doel is webtracking in de HTTP-header te identificeren. De aanpak bestaat uit drie delen: gegevensverzameling (deel 1), gegevensreductie (deel 2) en gegevensmodellering (deel 3). We voeren experimenten uit vanuit het perspectief van een eindgebruiker en leggen de onderzoeksgegevens vast. Hierna beschrijven we hoe we de big data filteren op de benodigde metadata. We laten zien dat *entropie* als rekeneenheid geen goed onderscheid maakt tussen tracking-cookies en cookies waarmee je niet gevolgd wordt. Met *small-data* elementen modelleren we de context van Europese digitale media. We werken achtereenvolgens een metadatamodel (een small-data model) uit en leggen de gemodelleerde gegevens vast in een graaf.

Hoofdstuk 4 bevat vier standpunten ten aanzien van RTB-systemen. We presenteren een *empirisch* standpunt van RTB als een netwerk van RTB-partners, die worden ingezet voor advertenties op nationale en regionale Europese nieuwswebsites. Vervolgens beschrijven we een *theoretisch* standpunt dat duidt op de belangrijkste concepten in RTB-partner netwerken. Hierna volgt een bespreking van een *juridisch* standpunt waarbij wordt ingegaan op verschillen in toestemmingsmechanismen voor tracking-cookies in EU-landen. Bovendien rangschikken we de landen op een schaal waarin de mate van het verkrijgen van toestemming tot uitdrukking komt. De conclusie is dat: (1) EU-landen die gedomineerd worden door een strikte implementatie van toestemmingsmechanismen zich bevinden aan de linkerkant van de schaal en (2) EU-landen waar impliciete toestemmingsmechanismen de overhand hebben zich aan de rechterkant van de schaal bevinden. Dit wordt ondersteund door de GraafgeBaseerde Methodologische Aanpak (GBMA) toe te passen met de network science al-

goritmen op de contextuele onderzoeksgegevens. Het onderzoek wordt afgesloten met een *maatschappelijke* standpunt. Onze belangrijkste bevinding is dat de rollen van de bedrijven die betrokken zijn bij het volgen van websitebezoekers uitgelegd kan worden door metadata uit de HTTP-header weer te geven in (wiskundige) kenmerken van de knooppunten en verbindingen van de graaf.

In hoofdstuk 5 wordt een antwoord gegeven op de probleemstelling en formuleren we de conclusies van het proefschrift. Daarna geef ik richtingen aan voor verder onderzoek.

Mijn minisamenvatting is derhalve als volgt. Dit proefschrift levert drie bijdragen aan de wetenschap. Allereerst, voor zover ik heb kunnen nagaan, is dit de eerste (empirische) studie waarbij een GraafgeBaseerde Methodologische Aanpak (GBMA) toegepast wordt op network science algoritmen voor het meten van de privacycomponent van RTB. Ten tweede wordt in het proefschrift een algemene methode beschreven voor het categoriseren van RTB-systemen op basis van cluster-edge-betweenness- en node-betweenness-algoritmen. Ten derde wordt de GBMA gedemonstreerd met behulp van een dataset van Europese nationale en regionale kranten.

Acknowledgments

A special word of thanks goes to Dr. Adriaan in 't Groen and Dr. Johannes Magliano-Tromp, who have been supportive from the moment I was welcomed at the Dual PhD Centre, Faculty of Governance and Global Affairs of Leiden University. The Dual PhD Centre offered a wonderful environment that cultivates the scientific prerequisites and social values necessary for this undertaking. Moreover, I would like to recognize two persons in particular for their support: Joke Hellemons and Petra Evers.

It is a pleasure to thank Wilbert Tomesen and Jacob Kohnstamm for supporting me throughout my journey. The completion of the support by the Autoriteit Persoonsgegevens was performed by the new chairman Aleid Wolfsen who immediately showed enthusiasm for the topic.

Then, I would like to express my deep gratitude to Jeff Chester. I also thank Udo Oelen, who made the thesis possible. I thank Aldo Cortesi who made the data collection possible by his technical support and Dr. Frank Takes for insightful comments. I would like to say a word of appreciation to Hannes Tschofenig and Rigo Wenning for introducing me to the standards community. Not to be forgotten are David van Altveer, Dr. Alexander Dix, Dr. Balachander Krishnamurthy, Sjoera Nas, Mike O'Neill, Dr. Petros Papadopoulos, Cecile Schut, and Floor Terra.

I thank the members of the assessment committee for the time taken to read the thesis. My supervisors can be thanked more explicitly since they have no say in the final assessment. Prof. dr. H. Jaap van den Herik and Dr. Mark Dechesne made true what they had promised and that was much.

I am grateful to my parents and my family. Last, and therefore most important I thank Delia Chin-Ten-San. I dedicate this thesis to you.

Publications and presentations

ORCID: <https://orcid.org/0000-0002-2203-6819>.

- IWGDPT (2014). Working Paper on Web Tracking and Privacy: Respect for context, transparency and control remain essential. In Dr. Alexander Dix (Ed.), *International Documents on Data Protection in Telecommunications and Media* (Vols. 1983–2013, pp. 492–514). Berlin: Berliner Beauftragter für Datenschutz und Informationsfreiheit. (Working paper adopted in 2013. First official draft by Van Eijk, R.)
- O'Neill, M., & Van Eijk, R. (2016). *Do Not Track 1.1 (Unofficial Draft)*. [Position Paper]. Retrieved 10 Januari 2017, from <https://github.com/w3c/dnt/blob/master/DNTBugs.html>
- Tschofenig, H., Leiba, B., Cook, B., & Van Eijk, R. (2011). Browser support for the Open Authorization (OAuth) Protocol. In *W3C workshop Identity in the Browser*. Mountain View, CA: W3C. [Position paper]. Retrieved 11 January 2014, from http://www.w3.org/2011/identity-ws/papers/idbrowser2011_submission_32.pdf
- Tschofenig, H., & Van Eijk, R. (2011). Do Not Track, An Attempt to Frame the Debate. In *W3C workshop Web Tracking and User Privacy*. Princeton, NJ: W3C. [Position paper]. Retrieved 11 January 2014, from <http://www.w3.org/2011/track-privacy/papers/Tschofenig.pdf>
- Van den Herik, H. J., & Van Eijk, R. (2013). Social Innovation & Big Data. In *IT Innovation Day*. Amersfoort: IDG Nederland. [Presentation].
- Van den Herik, H. J., & Van Eijk, R. (2014). The all-pervasive impact of Big Data in daily life: from Causality to Correlation. In *Launch of the Future Center - Smart Systems*. Leiden: Leiden Institute of Advanced Computer Science, Leiden University. [Presentation].
- Van den Herik, H. J., & Van Eijk, R. (2015). Big Data: een Introductie. In *Ondernemen en Big Data*. The Hague: VNO-NCW WEST. [Presentation].

- Van den Herik, H. J., & Van Eijk, R. (2016a). Big Data and Judicial Decisions. In *LERU Summer School 2016: Data Stewardship for Scientific Discovery and Innovation*. Leiden: League of European Research Universities. [Presentation]. doi: <http://dx.doi.org/10.2139/ssrn.2809612>
- Van den Herik, H. J., & Van Eijk, R. (2016b). De haakjes van Big Data. In *(Big) Data festival*. The Hague: Ministry of Infrastructure and Environment. [Presentation].
- Van Eijk, R. (2011). *Web Tracking Detection System (TDS): An effective strategy to reduce systematic monitoring and profiling of user habits across websites*. MSc thesis. Leiden: Leiden University, Leiden Institute of Advanced Computer Science. Retrieved 5 March 2018, from <http://liacs.leidenuniv.nl/assets/Masterscripties/ICTiB/Rob-van-Eijk-non-confidential.pdf>
- Van Eijk, R. (2012). *International Perspectives on Cloud Computing*. 36th Datenschutzfachtagung. [Presentation]. Köln: DAFTA.
- Van Eijk, R. (2013a). *Big Data: the Big Disconnect*. Workshop Formal methods for the informal world. [Presentation]. Leiden: Lorentz Center.
- Van Eijk, R. (2013b). *Forensic Aspects in Digital Investigations*. Colloquium. [Presentation]. Worcester, MA: Worcester Institute of Technology.
- Van Eijk, R. (2015). Towards a Delicate Balance in Strategic Innovations. In *Leadership challenges with big data. Module 10: Ethical, Legal and Privacy Challenges*. Rotterdam: Rotterdam School of Management, Erasmus University. [Presentation]. doi: <http://dx.doi.org/10.2139/ssrn.2696674>
- Van Eijk, R. (2016a). *A brief introduction to Real Time Bidding (RTB)*. [Presentation]. Brussels: Article 29 Working Party. doi: <http://dx.doi.org/10.2139/ssrn.2856941>
- Van Eijk, R. (2016b). Informatie: Leidt meer informatie tot minder vrijheid? In A. in 't Groen, C. De Roen, & P. Slaman (Eds.), *Leids-Haags Allegorieënlexicon* (pp. 151–160). Delft: Uitgeverij Eburon.
- Van Eijk, R. (2016c). Update on Privacy Legislation. In *Leadership challenges with big data. Module 10: Ethical, Legal and Privacy Challenges*. Rotterdam: Rotterdam School of Management, Erasmus University. [Presentation]. doi: <http://dx.doi.org/10.2139/ssrn.2804720>
- Van Eijk, R. (2017a). *On browser settings, cookies, and (not) being tracked by digital advertisements*. Working Party on Telecommuni-

- cations and Information Society. [Presentation]. Brussels: Council of the European Union. doi: <http://dx.doi.org/10.2139/ssrn.3003152>
- Van Eijk, R. (2017b). Tracking classification: a cluster analysis approach to identify missing values. In N. Bielova, N. Laoutaris, A. Narayanan, & N. Nikiforakis (Eds.), (Vol. 7, p. 103). Dagstuhl, Germany: Schloss Dagstuhl–Leibniz-Zentrum für Informatik. [Peer Reviewed Report]. doi: <http://dx.doi.org/10.4230/DagRep.7.4.83>
- Van Eijk, R. (2018a). Processing personal data: privacy requirements and security measures. In *Leadership challenges with big data & analytics*. Rotterdam: Rotterdam School of Management, Erasmus University. [Presentation]. doi: <http://dx.doi.org/10.2139/ssrn.3285144>
- Van Eijk, R. (2018b). *Technology for Privacy: Protecting Against Online Tracking and Profiling*. [Presentation]. Brussels: Computers, Privacy, and Data Protection congress. doi: <http://dx.doi.org/10.2139/ssrn.3106582>
- Van Eijk, R., & Toubiana, V. (2017). *Browser settings and consent: A brief introduction to Do Not Track (DNT)*. [Presentation]. Brussels: Committee on Civil Liberties, Justice and Home Affairs.
- Van Wissen, A., Kamphorst, B., & Van Eijk, R. (2013). A Constraint-Based Approach to Context. In P. Brézillon, P. Blackburn, & R. Dapoigny (Eds.), *Modeling and Using Context* (Vol. 8175, pp. 171–184). Heidelberg: Springer. [Peer Reviewed Journal].
- Wefers Bettink, W., Van Eijk, R., & Wagner, F. (2012). *Strictly Speaking: Cookies, Consent and Compliance*. Europe Data Protection congress. [Presentation]. Brussels: IAPP.

Curriculum Vitae

Robbert J. van Eijk (1969, Rosmalen) completed secondary school (1988, Gymnasium- β) at the Sint-Janslyceum in 's-Hertogenbosch. He received an undergraduate degree in electrical engineering (1990) from Eindhoven University of Technology, a BSc in electrical engineering (1994) from Fontys University of Applied Sciences in Eindhoven, and a MSc in ICT and business (2011) from the Leiden Institute of Advanced Computer Science, Faculty of Science of Leiden University.

Robbert works as senior supervision officer at the Autoriteit Persoonsgegevens, the Dutch Data Protection Authority (DPA). He represented the Dutch DPA in (inter)national meetings and as technical expert in court. He also represented the Article 29 Data Protection Working Party in the multi-stakeholder negotiations of the World Wide Web Consortium (W3C) on Do Not Track.

He started his professional career in the automotive industry. As an onsite consultant he specialized in dealer-network planning using geographical information systems. Then, he founded Blaeu Business Intelligence, a company with a focus on office automation for small-sized enterprises. The company was acquired after he had been running it successfully for nine years.

He worked together with SBV Forensics, a company specialized in financial forensic investigation. Together they assisted the official receiver after the bankruptcy of KPNQwest. They also collaborated in preparation of a court case after the bankruptcy of Ceteco.

Recently, he served as an external supervisor of the concluding Master's thesis projects of Scheepers [2018] and Turcios Rodríguez [2018]. He is an invited expert in the W3C Tracking Protection Working Group and the W3C Privacy Interest Group. He also contributes to the Inria PrivaWEB project on web tracking, e-Privacy Regulation, and user protection.

Index

NOTE: This index contains keywords representing concepts commonly used in Web Privacy Measurement (WPM) and Real-Time Bidding (RTB) technology. The keywords are listed alphabetically. In some cases they are extended to closely related (self-contained) key concepts.

B

bidding-as-a-service, 145

big data, 31–34

and internet platforms,
35

application of big data,
34

meta definition

analysis, 33

exchange, 33

interpretation, 33

maintenance, 33

real-time, 33

retention, 33

search, 33

variety, 32

velocity, 32

visualization, 33

volume, 32

narrative science, 33

bridging node, 187

C

C-steps, 90–92

central node, 187

D

data, 30

big data, 31–34

context, 5, 21, 28, 31, 63,
134, 146, 153, 179,
185, 193, 196,
203–204, 206

information, 31

knowledge, 31

metadata, 38–40

missing values, 124

small data, 36–38

Data Management Platform,
145

deal ID, 149

Demand Side Platform, 144

dendrogram, 161, 164

branches, 164

clades, 164

leaves, 164

phylogenetic tree, 165

display advertising

ad-slots, 142

publisher, 143

remnant inventory, 203

E

empirical view, 133–140,
153–171, 195

entropy, 63, 68, 85, 105, 111,
113–118
 high-entropy identifier,
 114
 Shannon entropy, 63,
 114
event tracking, 49–53

F

fingerprinting
 active, 60, 65–67
 browser extensions, 99
 cookie-like, 61
 cross-device, 61
 mitigations, 67
 passive, 60
 risks, 62
 stateful, 60
 stateless, 60
FourthParty, 72, 74, 83
FPDetective, 73, 83

G

graph, 106
 cluster-edge
 betweenness, 183
 Cypher, 107, 111, 122
 eigenvector centrality,
 190
 induced subgraph, 110
 isomorphism, 119
 node betweenness,
 184–185
 refactoring, 127, 157,
 184, 186–188, 266
 self-referencing loop,
 157, 166, 190
 subgraph, 108–110
 subgraph isomorphism,
 119
Graph-Based
 Methodological

Approach, 42, 79,
81, 102, 105, 129,
132, 180, 196, 197,
199, 200, 205

H

header bidding, 144
heatmap, 161
HyperText Transfer Protocol,
68
 header field
 registration, 70
 redirection, 121, 157
 request header, 69
 response header, 70
 standard, 71

I

identifier, 40
 cookie, 37
 false negative, 117
 false positive, 116
 Youtube-nocookie, 181,
 184
identity, 41
information, 31
interconnecting node, 188
inventory source, 142

K

knowledge, 31

L

legal view, 171–179, 195

M

M-steps, 129, 157–158
metadata, 38–40, 47
 difference between data
 and metadata, 38
 importance of metadata,
 39
 twenty common types,
 63

Mitmproxy, 73, 74, 86–87, 89,
96, 97, 103–104, 127
multidisciplinary
 legal-science
 paradigm, 55

N

narrative science, 33
Netograph, 77, 89, 154, 157
network of partners, 133,
195

O

OpenWPM, 74, 81, 83, 86,
117

P

privacy
 acquisitions, 76,
 133–134
 and webtracking, 21–30
 and WPM themes, 46
 commercial tracking, 48
 consent
 implementation
 differences, 175
 implied, 175
 strict, 174
 enforceability, 183, 204
 fundamental right, 22,
 24, 27, 202–205
 information asymmetry,
 53
 joint controller, 202
 mergers, 76, 133, 135
 partnerships, 76, 133,
 135–137, 195
 political advertising
 online, 48–49
 principles, 22–23
 state tracking, 48
 user control, 204

Westin, 22
profiling, 29, 98
prototyping, 54, 205

R

R-steps, 103–104
Real-Time Bidding
 bidding-as-a-service,
 145
 deal ID, 149
 definition, 20–21
 header bidding, 144
 missing values, 124
 node classification,
 124–128, 128
 node types, 185–190
 private deal, 148
 process, 150–152
 protocol, 146–148
 sequence diagram, 151
 system, 140
RTB-steps, 150–151, 154, 174,
190

S

satellite node, 187
small data, 36–38
societal view, 180–192, 196
Supply Side Platform, 143

T

Tag-Based Integration, 149
theoretical view, 140–152,
195
top-100 edges, 163, 167
top-20 edges, 161, 178

W

web privacy measurement,
45–49
 deep crawl, 91, 92,
 153–156

- long tail phenomenon,
 - 94–95
- research methodologies,
 - 54
- research themes, 46
- shallow crawl, 92, 117,
 - 154, 171
- visualization tools, 77
- web tracking, 25, 27–28
 - cross border, 160,
 - 168–170, 193
 - definition, 28, 106
 - deterministic
 - identification, 58
 - event tracking, 49–53
 - extended model,
 - 111–113, 118
 - generic model, 105–107,
 - 123, 124, 156
 - probabilistic
 - identification, 58
 - profiling, 29, 98
 - stateful, 61
 - stateless, 61
- WebXray, 76