



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