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

1

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 (herinafter: 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 incrementiality 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 of 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]

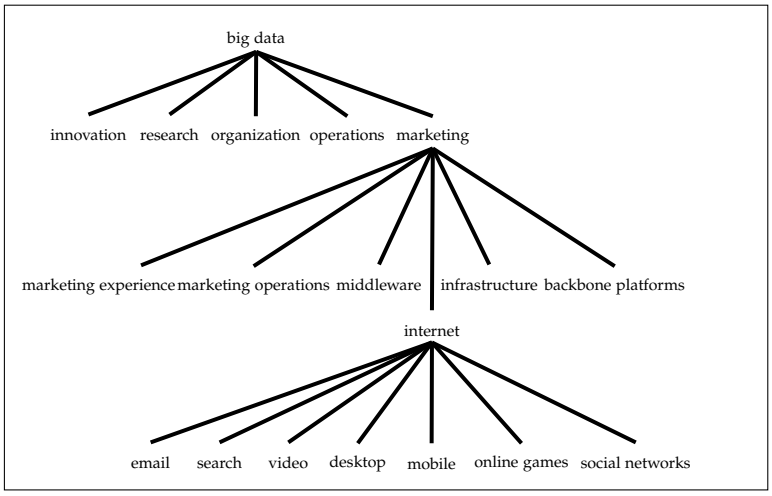

Figure 1.1: Big data and internet platforms. In the thesis we focus on internet applications of big data, i. e., desktop browsers and mobile browsers.

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 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 Rannenberg [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	✓			✓	✓

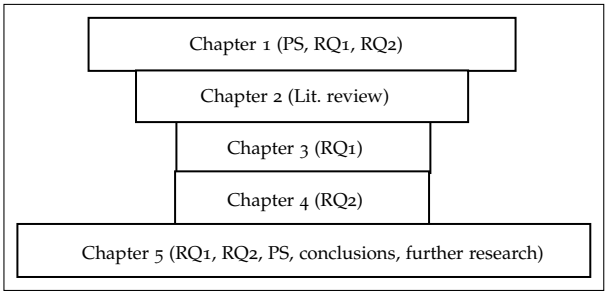

Figure 1.2: Visual structure of the thesis.

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.