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

Eijk, R.J. van

4

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/3pascertification/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]

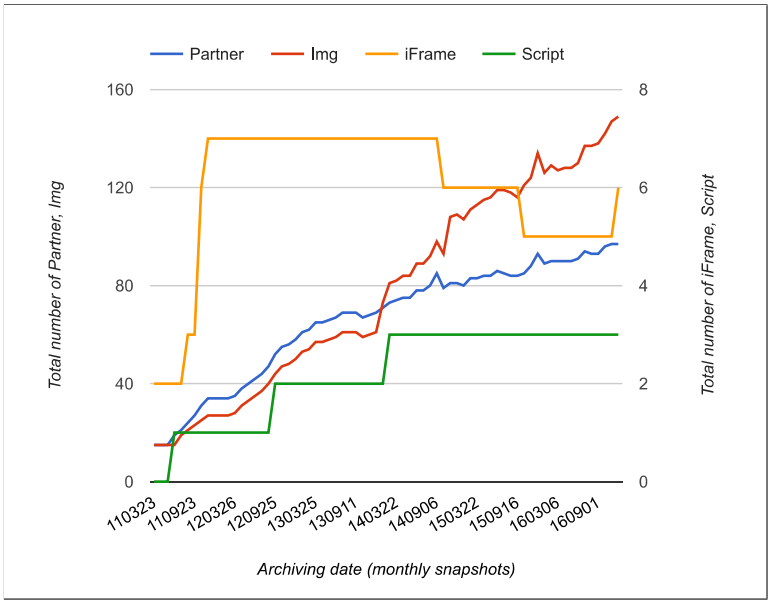

Figure 4.1: Longitudinal analysis of the Rubicon Project partner network between March 2011 and December 2016.

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).

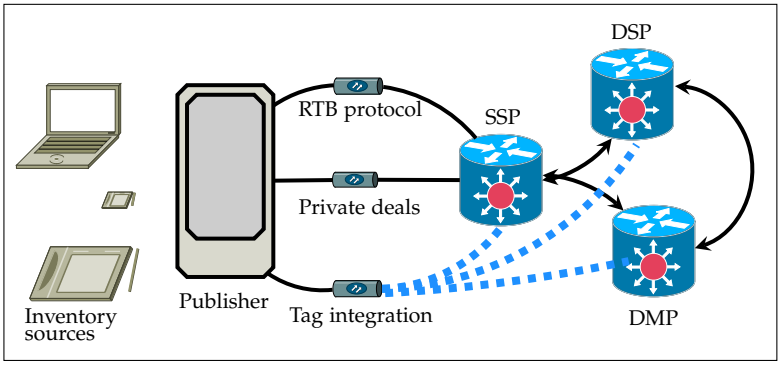

Figure 4.2: Theoretical view of RTB systems.

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: v0.10.3dev).

is leading with 54.7% followed by the leaderboard (29.7%) and the skyskraper (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 Wallmart 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.

```
POST /auctions HTTP/1.1
Content-Type: application/json
Content-Length: 640
accept: */*
connection: Keep-Alive
x-openrtb-version: 2.3
(...)
"user": {
        "id": "23456",
        "buyeruid": "202122",
        "data": [{
                "id": "303132",
                "name": "Data Provider X",
                "segment": [{"name": "online news"
                }, {
                        "id": "505148",
                        "name": "data_X_location",
                        "value": "Midwest USA"
                }, {
                        "id": "505152",
                        "name": "data_X_age",
                        "value": "40-50"
                }, {
                        "id": "404142",
                        "name": "data_X_buying_intent",
                        "value": "high" }]
                }]
        }
```

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 '0x0'-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*.

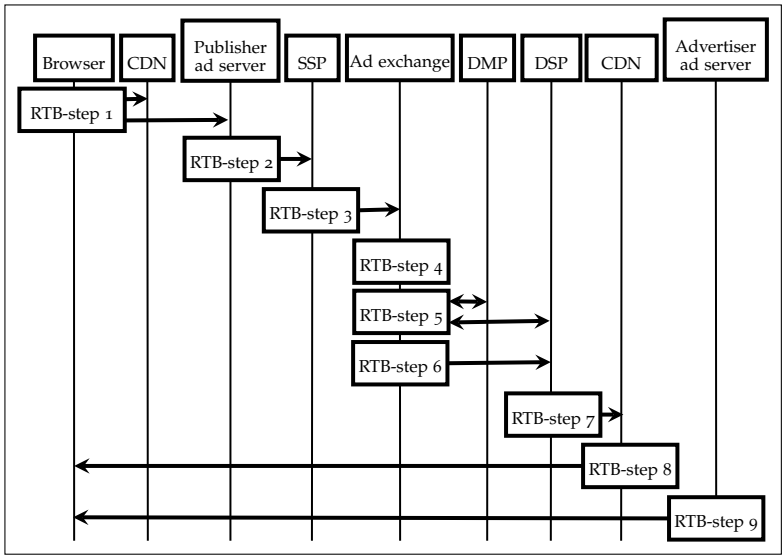

Figure 4.3: RTB sequence diagram with nine RTB-steps.

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.eufeeds.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 (Crawl7) 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, Crawl2, (...) Crawl6 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.philileftheros.com/` (29 August 2016).

302 In Luxembourg I crawled 19 news items from one newspaper: L'Essentiel, URL: `www.lessentiel.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.

```
(...)
"browser_initiator": {
  "linenumber": 119,
  "type": "parser",
  "url": "https://www.natuurlijkehaarkleuring.nl/afspraak/"
},
"browser_type": "Script",
"document_url": "https://www.natuurlijkehaarkleuring.nl/
    afspraak/",
"host": "d3gxy7nm8y4yjr.cloudfront.net",
"request": {
  "headers": {
    "Accept": "*/*",
    "Accept-Encoding": "gzip, deflate",
    "Connection": "keep-alive",
    "Host": "d3gxy7nm8y4yjr.cloudfront.net",
    "User-Agent": "Mozilla/5.0 (X11; Linux x86_64) AppleWebKit
        /537.36 (KHtml, like Gecko) HeadlessChrome
        /63.0.3239.132 Safari/537.36"
  },
  "method": "GET",
  "referrer_policy": "no-referrer",
  "url": "https://d3gxy7nm8y4yjr.cloudfront.net/js/embed.js"
},
(...)
```

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 county 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/talgalili/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.

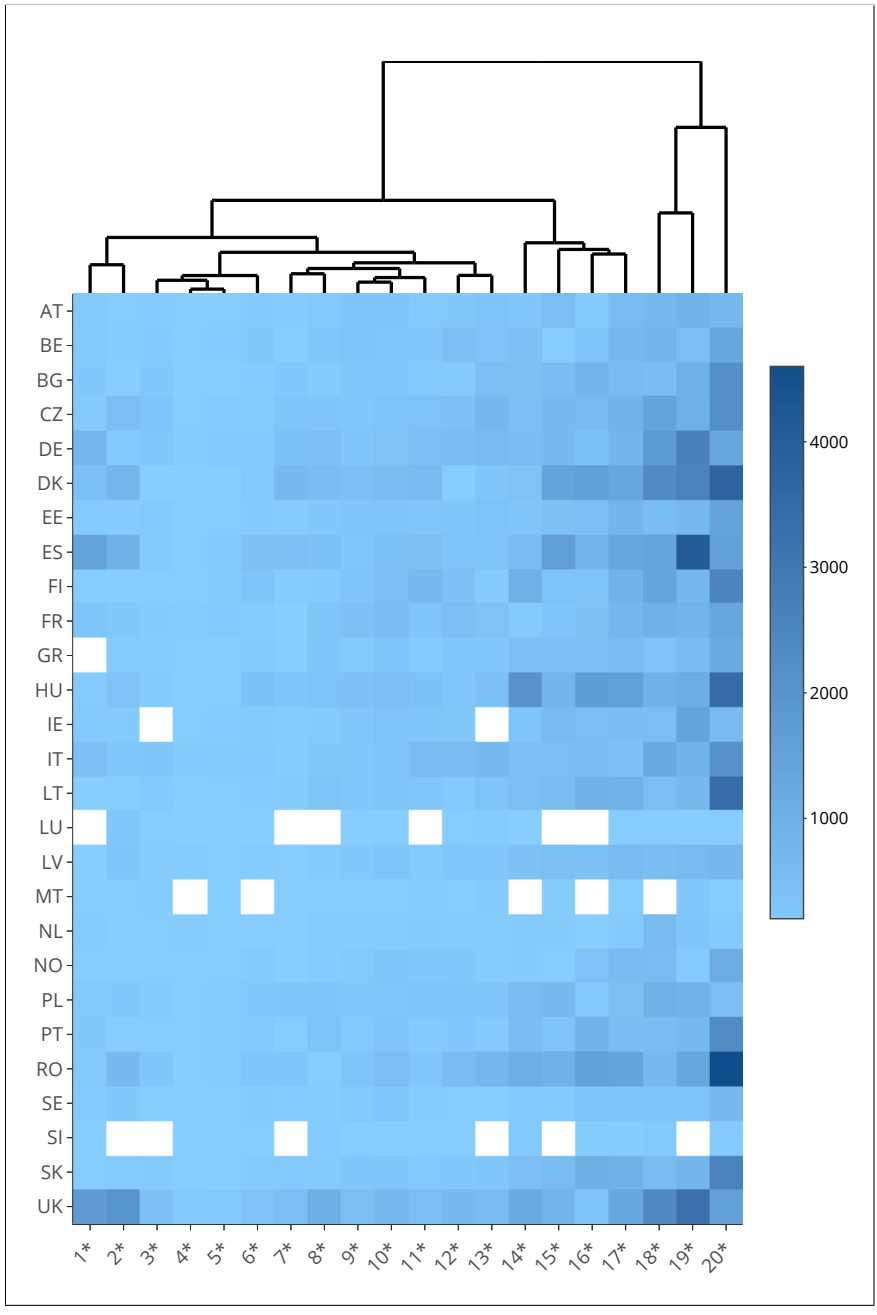

Figure 4.4: EU heatmap and dendrogram.

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.

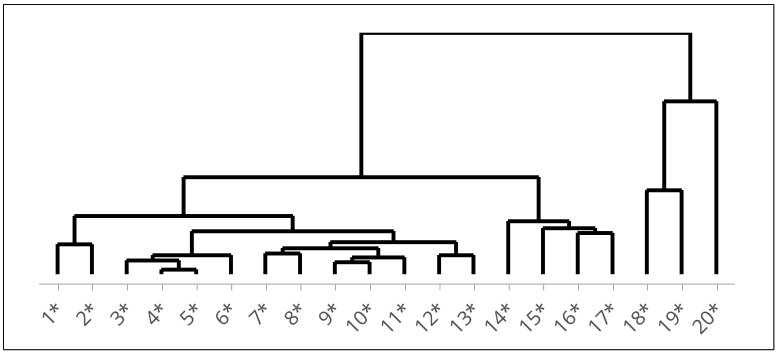

Figure 4.5: EU dendrogram.

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
PT (Portugal)	3,753
EE (Estonia)	1,908
IT (Italy)	1,687
AT (Austria)	1,468
NO (Norway)	1,397
IE (Ireland)	614
SE (Sweden)	564
PL (Poland)	454
DE (Germany)	369
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)

TYPES	TOTAL
Min.:	22
1st Qu.:	369
Median:	1,687
Mean:	4,355
3rd Qu.:	7,717
Max.:	18,741
Total:	108,875
Distinct edges:	252
Distinct nodes:	273

(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 county 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 GDPD - 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).

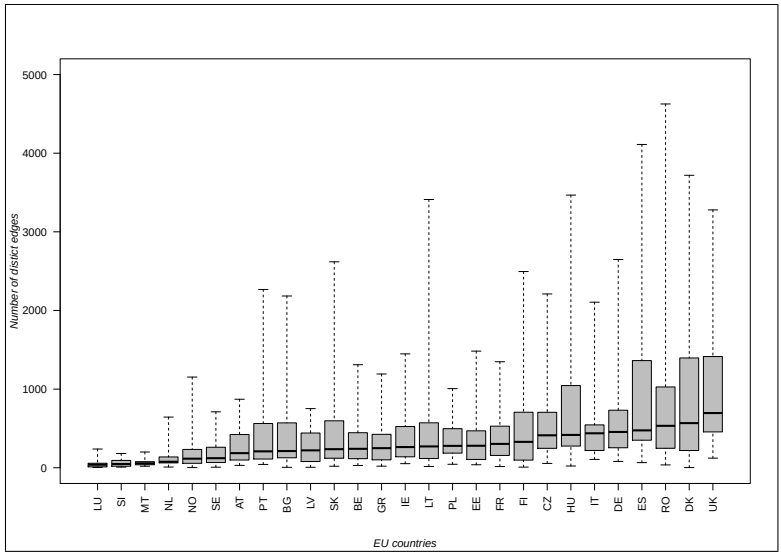

Figure 4.6: Edge boxplot of the subset top-20 edges. Cronbach's alpha: 0.963183 (sample size = 20, number of countries = 27).

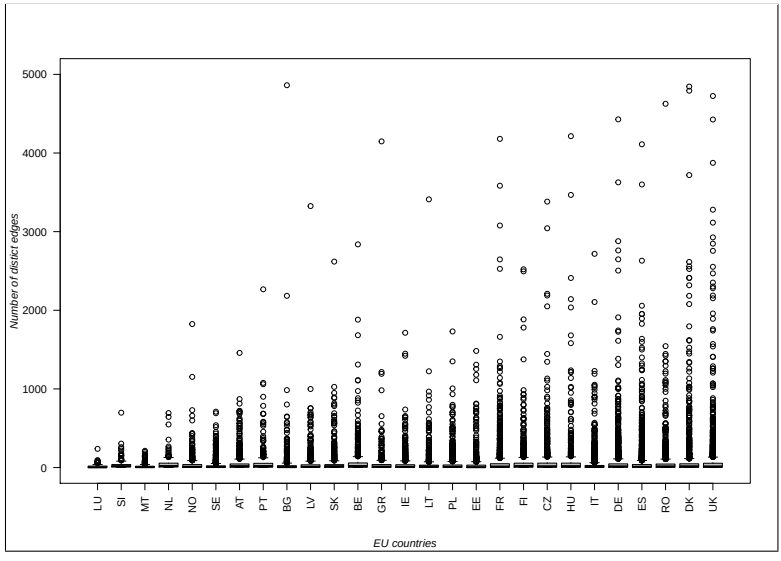

Figure 4.7: Edge boxplot of the full dataset. Cronbach's alpha: 0.8803147 (sample size = 13,789, number of countries = 27).

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","
         grey","grey","grey","grey","grey","grey","grey","grey","
         grey","grey","grey","grey","grey","grey","grey","grey","
         grey","grey","grey"),
8    at = c(1,2,3,4,5,6,7,8,9,10,11,12,13,14,15,16,17,18,
9    19,20,21,22,23,24,25,26,27), ylim=c(0,5000))
10 # Adding the x-axis and y-axis labels
11 mtext("EU countries", side = 1, line = 4, cex = 1, font = 3)
12 mtext("Distict 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

$$C_B(e) = \sum_{s \neq t} \frac{\sigma_{st}(e)}{\sigma_{st}} \tag{4.1}$$

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., Dnevik[338] and Mladina[339]. 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).

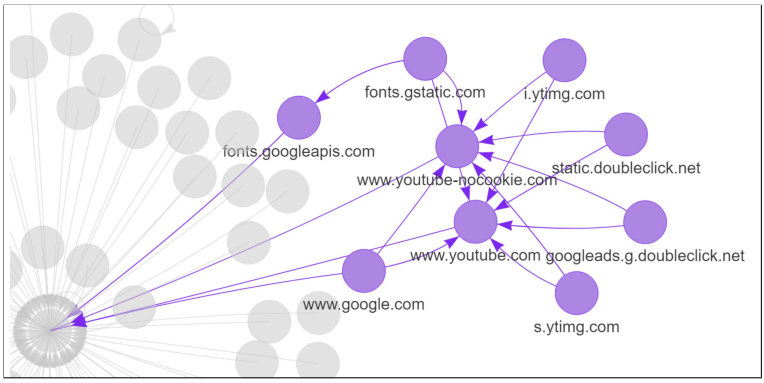

Figure 4.8: Cluster-edge betweenness: Slovenia (SI) with N=10 nodes and E=17 edges.

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 [Thieurmel, 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:

$$C_B(v) = \sum_{s \neq v \neq t \epsilon V} \frac{\sigma_{st}(v)}{\sigma_{st}} \tag{4.2}$$

where the sum runs over all s, t pairs with $s \neq t \neq v \epsilon V$.

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).

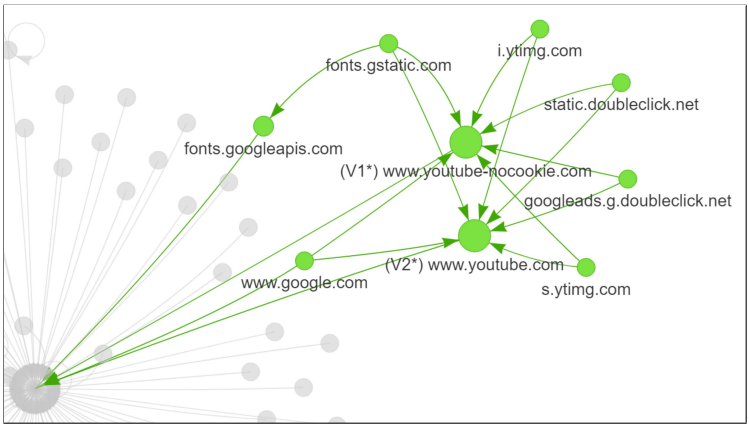

Figure 4.9: Central nodes: Slovenia (SI). Weighted nodes for the betweenness cluster YouTube with N=10 nodes and E=17 edges, g = (V, E), and two central nodes V_1*: 'www.youtube-nocookie.com', V_2*: 'www.youtube.com'.

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, Barthelemy, 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]

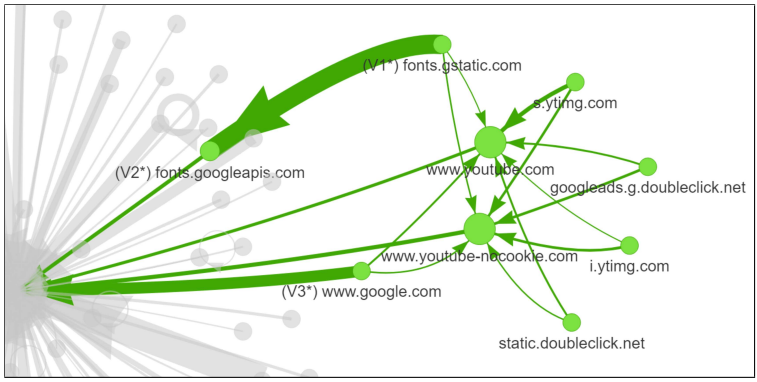

Figure 4.10: Annotated cluster: (Slovenia, SI). Weighted nodes and weighted edges for the betweenness cluster YouTube with N=10 nodes and E=17 edges, g = (V, E), E_1: V_1*xV_2* ('fonts.gstatic.com' -> 'fonts.googleapis.com'), and E_2: V_3*xV_{io}* ('www.google.com' -> 'crawled.io').

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*xV_2* (E_1) and V_3*xV_{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*xV_2* and E_2: V_3*xV_{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: Vc, Definition 4.16), satellite node (henceforth: Vs, Definition 4.17), bridging node (henceforth: Vb, Definition 4.18), and interconnecting node (henceforth: Vic, Definition 4.19). Moreover, we use Vio, 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 Vc_i with a high node-betweenness score in comparison to the remaining nodes in the same cluster $G(\alpha)$; with $Vc_i \epsilon V_\alpha$ and $Vc_i \neq V_{io}$.

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, $Vc(\alpha)$ representing one or more central nodes, and $Vb(\alpha)$ represents zero or more bridging nodes. A *satellite* node is defined to be a node Vs_i connected to $Vc(\alpha)$ but not to V_{io} by its shortest path ($Vs_i \epsilon V_\alpha$ with $Vs_i \neq V_{io}$, $Vs_i \neq Vc(\alpha)$, $Vs_i \neq Vb(\alpha)$).

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, $Vc(\alpha)$ representing one or more central nodes, and $Vs(\alpha)$ represents zero or more satellite node. A *bridging* node is defined to be a node Vs_i connected to V_{io} and Vc(i) by its shortest path ($Vb_i \epsilon V_\alpha$ with $Vb_i \neq V_{io}$, $Vb_i \neq Vc(\alpha)$, $Vb_i \neq Vs(\alpha)$).

DEFINITION 4.19: INTERCONNECTING NODE. Let us consider a refactored referrer graph denoted by g = (V, E) with multiple betweenness-centrality clusters $\Gamma(i) \epsilon G$, $G = ((V_\alpha, E_\alpha), (V_\beta, E_\beta), ..., (V_N, E_N))$, with $i \epsilon N$ 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 $E(\alpha)$: $V(\alpha)xVc(\beta)$.

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 Crqwl7 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	Vc_1	Vs_1–Vs_{11}	Vb_1–Vb_3	-
Twitter	Vc_2	Vs_{12}–Vs_{14}	-	-
Youtube	Vc_3	Vs_{15}–Vs_{18}	-	Vic_1
Spotify	Vc_4	Vs_{19}–Vs_{21}	Vb_4	-
Google	Vc_5	Vs_{22}	Vb_5, Vb_6	Vic_2
Doubleclick	Vc_6	Vs_{23}	-	-
Facebook	Vc_7	Vs_{24}	-	-
Cloudflare	Vc_8	-	Vb_7	-
'crawled.io'	Vio	Vs_{25}–Vs_{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

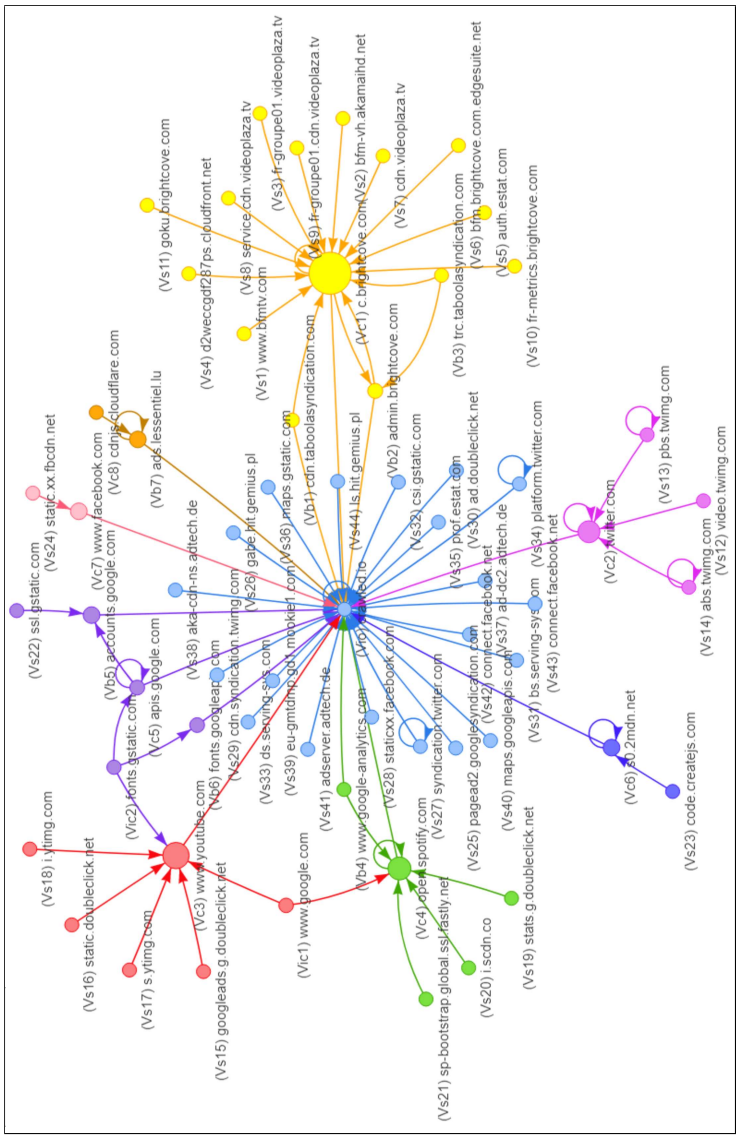

Figure 4.11: Annotated cluster: Luxembourg (LU) with N=62 nodes, E=81 edges, and C_B=9 clusters.

column 'interconnecting node'. The values in the table, e. g., Vc_1, Vs_1, Vb_1, or Vic_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 marketeers 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 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.

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?