Universiteit
Leiden
The Netherlands

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

Eijk, R.J. van

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

Version:	Not Applicable (or Unknown)
License:	Licence agreement concerning inclusion of doctoral thesis in the Institutional Repository of the University of Leiden
Downloaded from:	https://hdl.handle.net/1887/68261

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

Cover Page

Universiteit Leiden

The following handle holds various files of this Leiden University dissertation:
http://hdl.handle.net/1887/68261

Author: Eijk, R.J. van
Title: Web privacy measurement in real-time bidding systems. A graph-based approach to RTB system classification
Issue Date: 2019-01-29

Summary

Web Privacy Measurement in Real-Time Bidding System
A graph-based approach to RTB system classification

Web Privacy Measurement (WPM) is the observation of websites and services to detect, characterize, and quantify privacy impacting behaviors [Englehardt et al., 2014]. Real-Time Bidding (RTB) is a big data application within the organizational field of marketing to improve sales by real-time data-driven marketing and personalized (behavioral) advertising. The practice of RTB manifests itself on, e. g., news websites.

The aim of the thesis is to investigate and understand to what extent we can measure the privacy component connected to RTB. In particular, we focus on two related research questions: (1) how do we move from data collection to graph analysis? and (2) what are the emerging characteristics of the graph which are fit for graph analysis? Below we provide first a broad overview of the thesis and then we give a brief chapter-by-chapter description of the contents.

For our investigations, we crawled articles from European news websites with the help of the Netograph experimental framework. To collect specific research data in this context we adopted the assumption that today online news continues to rely on advertising as a business model.

Next, we filtered metadata from the HTTP headers of the crawl data. Then we built a small-data model with metadata derived from the web crawls. The small-data model expresses the interrelationships between RTB actors. We stored the network of RTB partners in a separate graph database and did so by EU country. We then refactored the graph databases by grouping the origins (domain names) of all the news items that we crawled on one single node. Subsequently, we applied two types of network science algorithms to our databases, i. e., (1) cluster-edge betweenness and (2) node betweenness.

First, cluster-edge betweenness turned out to be a strong indicator for the identification of the main actor(s) in a partner network.

Second, node betweenness showed to be a strong indicator for networks of RTB partners. The application of the two algorithms enabled us to categorize the nodes (i. e., actor(s) in an RTB-partner network) into four types:

(1) bridging node,
(2) interconnecting node,
(3) central node, and
(4) satellite node.

The first two types are the result of the application of cluster-edge betweenness. The third and fourth type are the result of the application of node betweenness.

Based on the experiments we arrived at three conclusions. First, cluster-edge betweenness and node betweenness help us in understanding the interconnectedness of the ad-technology companies. Second, we may conclude that the *combined* application of both network science algorithms to empirical WPM-data enables us to categorize a partner network. Third, the research also shows that the interconnection in a betweenness cluster is caused by the data flows of the companies themselves due to their specializations in ad technology. This is the main reason why categorization of RTB-partner networks with cluster-edge betweenness and node betweenness algorithms is effective.

In Chapter 1, we provide a background for our investigation emphasizing the topics privacy and RTB systems. We start by briefly looking at data, big data, small data, and metadata. Then we provide our definition of web tracking and justify it by arguments. Our definition of web tracking is meant to position the thesis within the larger privacy debate. We present our problem statement and formulate two research questions. Furthermore, we delineate the research methodology to answer the research questions.

In Chapter 2, we review literature on Web Privacy Measurement (WPM). We do so by providing a description of (1) three client identification mechanisms, (2) deterministic and probabilistic identification, and (3) web-based fingerprinting. The descriptions serve as a background for the state of the art in web tracking. Furthermore, we briefly describe HTTP header fields. We close the chapter by providing an overview of (1) four different approaches to data collection in the academic field of WPM and (2) nine WPM visualization tools.

In Chapter 3, we develop our novel Graph-Based Methodological Approach (GBMA) to identify web tracking in HTTP header fields (flow data). Our approach consists of three parts: data collection (Part 1), data reduction (Part 2), and data modeling (Part 3). We conduct experiments from an end-user perspective. Of course, we retain the research data. For this purpose, we describe how we reduced the big data in such a way that we can perform our experiments with metadata from the HTTP header. We show that *entropy* as a measure fails to differentiate between tracking cookies and non-tracking cookies. Therefore, we model the context of online media with specific small-data elements. Then we provide a metadata model (a small-data model) and store the data into a graph.

In Chapter 4, we present four views on RTB systems. First, we present an *empirical* view of RTB as a network of partners deployed on national and regional European news websites. Then, we give a *theoretical* view of the key concepts in RTB systems reflected in partner networks. Thereafter, we discuss a *legal* view. Here we identify differences in consent implementations for tracking cookies in EU countries. To do so, we rank the countries on a scale. From the results we arrive at two conclusions: (1) EU countries are dominated by strict-consent mechanisms on the left end of the scale and (2) EU countries where implied-consent mechanisms have the upper hand are to be found on the right end of the scale. These two conclusions are supported by applying the GBMA with the network science algorithms to our contextual research data. Finally we completed our research by a *societal* view. Our main finding is that the roles of the companies involved with web tracking can be explained through expressing metadata from the HTTP header into (mathematical) characteristics of nodes and edges of the graph.

In Chapter 5 we provide an answer to the problem statement and there we formulate the conclusions of the thesis. Then we recommend directions for further research.

In summary, this thesis makes the following three contributions. First, to the best of our knowledge, it is the first (empirical) study to apply a Graph-Based Methodological Approach (GBMA) to network science algorithms for measuring the privacy component of RTB. Second, we describe a general method of categorizing RTB systems based on cluster-edge betweenness and node-betweenness

algorithms. Third, we demonstrate our GBMA (approach) on a dataset of European national and regional news websites.