



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

Samenvatting (Summary in Dutch)

WEB PRIVACY MEASUREMENT IN REAL-TIME BIDDING SYSTEMEN EEN GRAAFGEBASEERDE AANTAK VOOR DE CLASSIFICATIE VAN RTB SYSTEMEN

Web Privacy Measurement (WPM) is het observeren van websites en services met als doel om (mogelijke) inbreuken op de privacy (rechten en vrijheden) van betrokkenen te detecteren, karakteriseren en kwantificeren (cf. [Englehardt et al., 2014]). Real-Time Bidding (RTB) is een *big-data* toepassing binnen de marketingactiviteiten van een organisatie met als doel om de verkoop te verbeteren door middel van *real-time* datagedreven marketing en gepersonaliseerde (gedrags)advertenties. De praktijk van RTB manifesteert zich bijvoorbeeld op websites van Europese kranten en andere digitale media.

Het doel van mijn proefschrift is om te onderzoeken in hoeverre we de privacycomponent die verbonden is met RTB kunnen meten. We richten ons daarbij met name op twee onderzoeksvragen: (1) hoe gaan we van een dataverzameling naar graafanalyse? en (2) wat zijn de emergente karakteristieken van de graaf die geschikt zijn voor graafanalyses?

Voor ons onderzoek zijn nieuwsartikelen bezocht op Europese krantenwebsites op een geautomatiseerde wijze met behulp van de Netograph-webcrawldienst. Het volledige netwerkverkeer tussen de browser en de website is daarbij vastgelegd. De belangrijkste reden voor het verzamelen van onderzoeksgegevens in deze context is gebaseerd op de aanname dat het businessmodel van krantenwebsites nog steeds (sterk) afhankelijk is van inkomsten uit advertenties.

Vervolgens zijn HTTP-headers als metadata uit de ruwe data gefilterd. Hiermee kan een small-data model (dat is een gerichte subgraaf) gemaakt worden. Het small-data model drukt de onderlinge relaties uit tussen RTB-actoren, die in een afzonderlijke grafendatabase per EU-land zijn opgeslagen. De grafendatabase is geherstructureerd door het oorsprongpunten (domeinnamen) van alle digitale nieuwsitems die zijn bezocht te groeperen op

één enkel knooppunt. Daarbij zijn twee soorten network science algoritmen toegepast op onze database te weten (1) cluster-edge-betweenness en (2) node-betweenness.

In de context van RTB blijkt cluster-edge-betweenness een sterke indicator te zijn voor de identificatie van de belangrijkste actor(en) in een RTB-partnernetwerk. Node-betweenness blijkt een sterke indicator te zijn voor netwerken van RTB-partners. Door toepassing van de twee algoritmen kunnen we de knooppunten (d.w.z. de actor(en) in een RTB-partnernetwerk) categoriseren in vier typen:

- (1) overbruggingsknooppunt,
- (2) verbindend knooppunt,
- (3) centraal knooppunt en
- (4) satellietknooppunt.

De eerste twee typen zijn het resultaat van de toepassing van cluster-edge-betweenness. Het derde en vierde type zijn het resultaat van de toepassing van node-betweenness.

Uit de experimenten mogen we drie conclusies trekken. Ten eerste, cluster-edge-betweenness- en node-betweenness-algoritmen stellen ons in staat om de onderlinge verbondenheid van de advertentietechnologiebedrijven beter te begrijpen. Ten tweede, de gecombineerde toepassing van zowel network science algoritmen als empirische WPM-gegevens stellen ons in staat om een netwerk van RTB-partners te categoriseren. Ten derde toont het onderzoek ook aan dat de interconnectie in een betweenness-cluster veroorzaakt wordt door de datastromen van de bedrijven zelf vanwege hun specialisaties in advertentietechnologie. Dit is de belangrijkste reden waarom het categoriseren van netwerken van RTB-partners met cluster-edge-betweenness- en node-betweenness-algoritmen effectief is.

In hoofdstuk 1 wordt een achtergrond geschetst van het onderzoek in privacy- en RTB-systemen. Het verhaal begint met een kort overzicht van de begrippen gegevens: big data, small data en metadata. Tevens wordt een definitie gegeven voor de term webtracking (het gevolgd worden op internet). We motiveren deze concepten met argumenten. Vervolgens beoog ik met de definitie voor webtracking het proefschrift in een breder debat te plaatsen over privacy. Ik presenteer de probleemstelling en formuleer twee onderzoeksvragen. Verder wordt de onderzoeksmethodologie uiteen gezet.

In hoofdstuk 2 bespreken we de literatuur over Web Privacy Measurement (WPM). We geven een beschrijving van (1) drie client-identificatiemechanismen, (2) deterministische en probabilistische identificatie en (3) webgebaseerde vingerafdrukken. Deze beschrijvingen dienen als achtergrond voor de stand van zaken in webtracking, gevolgd door een korte beschrijving van een aantal velden in de HTTP-header. We sluiten het hoofdstuk af met een overzicht van (1) vier verschillende benaderingen van dataverzameling binnen het vakgebied van WPM en (2) negen visualisatiehulpmiddelen van WPM.

In hoofdstuk 3 staat de ontwikkeling van onze nieuwe GraafgeBaseerde Methodologische Aanpak GBMA centraal. Het doel is webtracking in de HTTP-header te identificeren. De aanpak bestaat uit drie delen: gegevensverzameling (deel 1), gegevensreductie (deel 2) en gegevensmodellering (deel 3). We voeren experimenten uit vanuit het perspectief van een eindgebruiker en leggen de onderzoeksgegevens vast. Hierna beschrijven we hoe we de big data filteren op de benodigde metadata. We laten zien dat *entropie* als rekeneenheid geen goed onderscheid maakt tussen tracking-cookies en cookies waarmee je niet gevolgd wordt. Met *small-data* elementen modelleren we de context van Europese digitale media. We werken achtereenvolgens een metadatamodel (een small-data model) uit en leggen de gemodelleerde gegevens vast in een graaf.

Hoofdstuk 4 bevat vier standpunten ten aanzien van RTB-systemen. We presenteren een *empirisch* standpunt van RTB als een netwerk van RTB-partners, die worden ingezet voor advertenties op nationale en regionale Europese nieuwswebsites. Vervolgens beschrijven we een *theoretisch* standpunt dat duidt op de belangrijkste concepten in RTB-partner netwerken. Hierna volgt een bespreking van een *juridisch* standpunt waarbij wordt ingegaan op verschillen in toestemmingsmechanismen voor tracking-cookies in EU-landen. Bovendien rangschikken we de landen op een schaal waarin de mate van het verkrijgen van toestemming tot uitdrukking komt. De conclusie is dat: (1) EU-landen die gedomineerd worden door een strikte implementatie van toestemmingsmechanismen zich bevinden aan de linkerkant van de schaal en (2) EU-landen waar impliciete toestemmingsmechanismen de overhand hebben zich aan de rechterkant van de schaal bevinden. Dit wordt ondersteund door de GraafgeBaseerde Methodologische Aanpak (GBMA) toe te passen met de network science al-

goritmen op de contextuele onderzoeksgegevens. Het onderzoek wordt afgesloten met een *maatschappelijke* standpunt. Onze belangrijkste bevinding is dat de rollen van de bedrijven die betrokken zijn bij het volgen van websitebezoekers uitgelegd kan worden door metadata uit de HTTP-header weer te geven in (wiskundige) kenmerken van de knooppunten en verbindingen van de graaf.

In hoofdstuk 5 wordt een antwoord gegeven op de probleemstelling en formuleren we de conclusies van het proefschrift. Daarna geef ik richtingen aan voor verder onderzoek.

Mijn minisamenvatting is derhalve als volgt. Dit proefschrift levert drie bijdragen aan de wetenschap. Allereerst, voor zover ik heb kunnen nagaan, is dit de eerste (empirische) studie waarbij een GraafgeBaseerde Methodologische Aanpak (GBMA) toegepast wordt op network science algoritmen voor het meten van de privacycomponent van RTB. Ten tweede wordt in het proefschrift een algemene methode beschreven voor het categoriseren van RTB-systemen op basis van cluster-edge-betweenness- en node-betweenness-algoritmen. Ten derde wordt de GBMA gedemonstreerd met behulp van een dataset van Europese nationale en regionale kranten.