Data-Driven Transparency About Online Tracking

Euirim Choi, Claire Dolin, Aaron Goldman, Chang Min Hahn, Shawn Shan, Ben Weinshel
University of Chicago
{euirim, cdolin, argold, changhahn, shansixioing, weinshel}@uchicago.edu

Michelle L. Mazurek
University of Maryland, College Park
mmazurek@cs.umd.edu

Blase Ur
University of Chicago
blase@uchicago.edu

Abstract
Targeting advertisements to specific users based on their browsing activity can be helpful for both users and advertising networks, yet many users also find this practice unsettling and privacy-invasive. Although a number of privacy tools can help users control tracking, average users are left utterly confused about online behavioral advertising (OBA) even after using such tools. We are working to move beyond existing tools, which alert users to tracking occurring at the current moment, by designing and testing a tool that takes a data-driven, personalized approach to privacy awareness. We describe our work in progress designing a browser extension that enables users to explore what information third-party companies have tracked about them over time, as well as what those companies may have inferred about their interests from this data. We are currently exploring the impact of presenting different abstractions and granularities of the information tracked, as well as evaluating user reactions and concerns related to different methods of making inferences and targeting ads.

Author Keywords
tracking; transparency; online behavioral advertising (OBA); data-driven; access; intelligibility

Introduction

In recent years, data-driven algorithms have begun to supplant humans in making many important decisions and classifications [17]. The processing capabilities of computers, along with the vast amount of data that can be easily collected, enable these algorithms to find hitherto hidden statistical correlations in data. As these algorithms make decisions or classifications about humans, individuals are often left completely unaware of why they were classified under a certain label. That is, the algorithms' decision-making processes are often opaque and completely unintelligible to humans [6, 9, 20, 21].

Vast amounts of personal data are commonly leveraged for algorithmic decision making in the course of daily internet usage, where users knowingly and unknowingly share huge amounts of personal data. Much of this data is collected automatically. Users' web queries, the links they click on, their geographic location while using electronic devices, and other behavioral metrics are recorded. This "big data" is fed into opaque inferencing algorithms that infer the user's demographics, preferences, and habits. These inferred attributes eventually modify that user's view of the web, including their search results, the ads they see, and even the prices they are quoted for products and services.

Many studies show that people say they are, to varying degrees, uncomfortable with the collection and tracking of their online activities [11, 13, 14, 19, 20, 24]. A number of existing browser extensions (e.g., Ghostery [2], Lightbeam [3], and Privacy Badger [4]) help users control how their data is tracked and used; however, these tools are not well-understood [15]. There are many interrelated reasons why users do not act more forcefully to prevent the collection of their behavioral data: because they appreciate the benefits of a personalized web [6, 20]; because tools to control tracking are so difficult and confusing to use [12]; because they do not believe they can prevent tracking [18]; and perhaps most importantly, because they simply do not understand the frequency, mechanics, and potential consequences of tracking [14, 20, 21].

To a limited extent, others have proposed and prototyped tools to integrate a sum of a user's behavioral information into explanations of online tracking [1, 5, 7, 8, 16, 23]. However, several of these tools use aggregate, rather than personal, data. Other tools use static visualizations, textual tables or complicated network graphs that are hard to interpret. The efficacy of most of these tools has been sparsely evaluated, though limited evaluations have found the tools contribute minimally to users' understanding of online tracking [15]. While we start from a similar premise, we will improve on these efforts, building a tool that will allow us to better understand how users respond to more information about: (1) personalized, longitudinal history data, (2) inferencing data that estimates not just who is tracking a user, but what items can be inferred about the user as a result, and (3) interfaces that support user-guided questions to further ensure our tools intuitively provide the information most critical to users.

The Tool

Goal: Help users understand data collection and inferencing using their personal browsing histories.
Expected outcomes: A tool that helps casual users understand web tracking.

The tool we are building will provide visualizations designed to help users understand how their personal activities are tracked around the web. Our tool will "track the tracking" that occurs as an individual user browses the web, storing *which companies* have tracked the user on *what* websites

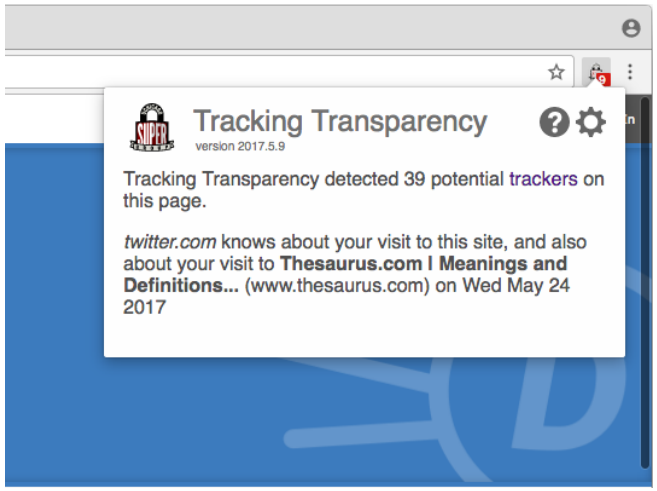

Figure 1: Our current plugin provides an example of longitudinal tracking on the current page.

(and *when*) in a local database on the user's computer. When the user encounters these third-party trackers in future browsing, the tool will provide personalized, longitudinal information about what trackers could know about the user's browsing history and inferences they could make regarding the user's interests. We hypothesize that by basing explanations of tracking around a user's own, personal examples [10], we can mitigate the great difficulties end users have understanding third-party online tracking [15, 20].

Our current prototype, adapted from the Electronic Frontier Foundation's Privacy Badger add-on [4], collects information about each page visit and the trackers on that page. A popup in the toolbar displays information about a specific tracker on the current page and a page visit on a different domain where the tracker was also present (Figure 1).

We hypothesize that showing users our best guesses about the inferences advertisers have made about their interests will help users understand how their data is used, a critical step in privacy awareness [22]. In fact, in initial pilot studies, more than two-thirds of participants agreed that a plugin showing the inferences companies had made about them would be useful. In Figure 2, we show one mockup of how inferences might be conveyed to users.

Inferences based on user browser activity are made as follows. We associate relevant Wikipedia articles to each of Google's ad interest categories. Then we use a modified graph-based TextRank algorithm to extract key unigrams (keywords) from the Wikipedia articles associated with each interest category. When a user visits a web page, we compare the words on the page to the keywords of each ad interest category. The category with the most matching keywords above a certain threshold is identified as the appropriate classification for the page. Through this comparison, we aim to identify and show to users the sorts of inferences that might be intuitive (e.g., browsing pages about Paris leads to subsequent ads for Parisian hotels), as well as some inferences that may not be intuitive.

Methodology

While we build the browser extension, we are also working on a user study to measure the usefulness and educational value of different possible UI designs through an online survey with two parts. In the first part, we show a number of potential "hooks," or attention-grabbing taglines, for the extension popup (Figure 1). We hope to gather information about how mentioning specific pages (e.g. "DoubleClick knows about your visits to '27 Best Things to Do in New York City,' 'Buy Tickets | The Metropolitan Museum of Art,' and 'The Ultimate NYC Ice Cream Shop Bucket List.'"), or mentioning long-term activity (e.g. "DoubleClick tracked

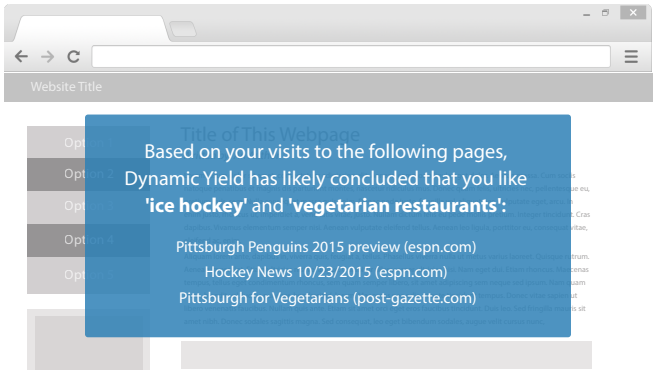

Figure 2: In this mockup visualization, a user can see examples of features tracking companies have inferred about them, along with the source of each inference.

you on the 47 amazon.com pages you visited in the past 30 days") changes users' perceptions about trackers.

In the second part, participants see a browser window with an ad that is relevant to their personal online shopping behaviors. Each user is given 11 explanations for why the ad is being targeted to them, in a randomized order. We are collecting quantitative data about which ad targeting methods users consider useful, which they are most comfortable with, and how important they find transparency about ad targeting. We will use the results from both parts of this survey to inform our choices about what information and inferences to prioritize in the extension display.

In the near future, we will run a field trial with our tracking tool. We will evaluate the impact of this tool on real users with real data, via a 75-participant, 2-week field trial. Participants will complete an entry survey (over the internet) assessing their prior knowledge of and attitudes toward inferencing, and then install our plugin on their own computer. Each will be assigned round-robin to one of several conditions, designed to compare different modalities for learning about tracking and inferencing. Potential conditions include:

1. Control condition: The participant will watch a short video about tracking and inferencing at the beginning of the study and use a plugin that provides a generic informational pop-up about tracking.
2. Longitudinal tracking without inferencing: The plugin will provide examples and visualizations of how each third party connects the websites a participant has visited previously.
3. Longitudinal tracking with inferencing: The plugin will also estimate the interests an advertiser or data broker might have inferred based on the user's browsing history. The plugin will present information showing these inferences.

After two weeks of using the specified plugin for everyday browsing, participants will complete an exit survey measuring changes in their knowledge of targeting and inferencing; their attitude toward online tracking and advertising, including their desire to block such tracking even if it reduces the relevance of ads they receive; and perceptions of the plugin's utility. The plugin will also automatically collect information about users' interactions with the different interface elements. To protect participants' privacy, information collected by the plugin will be aggregated and/or hashed as appropriate; detailed browsing data will not be collected.

Acknowledgments

This work is supported in part by a grant from the Data Transparency Lab (DTL) and by a UMIACS contract under the partnership between the University of Maryland and the DoD.

REFERENCES

1. 2017. Collusion. (2017). http://collusion.toolness.org.
2. 2017. Ghostery. (2017). http://www.ghostery.com http://www.ghostery.com.
3. 2017a. Lightbeam. (2017). http://www.mozilla.org/en-US/lightbeam/.
4. 2017. Privacy Badger. (2017). http://www.eff.org/privacybadger.
5. 2017b. Visualizing Lightbeam. (2017). http://research.ecuad.ca/lightbeam.
6. Lalit Agarwal, Nisheeth Shrivastava, Sharad Jaiswal, and Saurabh Panjwani. 2013. Do not embarrass: Re-examining user concerns for online tracking and advertising. In *Proc SOUPS*.
7. Julio Angulo, Simone Fischer-Hübner, Tobias Pulls, and Erik Wästlund. 2015. Usable Transparency with the Data Track: A Tool for Visualizing Data Disclosures. In *Proc. CHI Extended Abstracts*.
8. Yngvil Beyer, Erik Borra, Carolin Gerlitz, Anne Helmond, Koen Martens, Simeona Petkova, J C Plantin, Bernhard Rieder, Lonneke van der Velden, and Esther Weltevrede. 2012. Track the Trackers. (2012). https://wiki.digitalmethods.net/Dmi/DmiWinterSchool2012TrackingTheTrackers
9. Simone Fischer-Hübner, Julio Angulo, Farzaneh Karegar, and Tobias Pulls. 2016. Transparency, Privacy and Trust Technology for Tracking and Controlling My Data Disclosures: Does This Work?. In *Proc. IFIP International Conference on Trust Management*.
10. Marian Harbach, Markus Hettig, Susanne Weber, and Matthew Smith. 2014. Using personal examples to improve risk communication for security & privacy decisions. In *Proc. CHI*.
11. Chris Hoofnagle, Ashkan Soltani, Nathan Good, Dietrich Wambach, and Mika Ayenson. 2012. Behavioral Advertising: The Offer You Cannot Refuse. *Harvard Law & Policy Review* 6, 2 (2012), 273–296.
12. Pedro G. Leon, Blase Ur, Richard Shay, Yang Wang, Rebecca Balebako, and Lorrie Faith Cranor. 2012. Why Johnny can't opt out: A usability evaluation of tools to limit online behavioral advertising. In *Proc. CHI*.
13. Jonathan R. Mayer and John C. Mitchell. 2012. Third-party web tracking: Policy and technology. In *Proc. IEEE S&P*.
14. Aleecia M McDonald and Lorrie Faith Cranor. 2010. Americans' attitudes about internet behavioral advertising practices. In *Proc. WPES*.
15. Florian Schaub, Aditya Marella, Pranshu Kalvani, Blase Ur, Chao Pan, Emily Forney, and Lorrie Faith Cranor. 2016. Watching Them Watching Me: Browser Extensions' Impact on User Privacy Awareness and Concern. In *Proc. USEC*.
16. Yuuki Takano, Satoshi Ohta, Takeshi Takahashi, Ruo Ando, and Tomoya Inoue. 2014. MindYourPrivacy: Design and implementation of a visualization system for third-party Web tracking. In *Proc. IEEE PST*.
17. Omer Tene and Jules Polonetsky. 2013. Big data for all: Privacy and user control in the age of analytics. *Nw. J. Tech. & Intell. Prop.* 11, 5 (2013).

18. Joseph Turow, Michael Hennessy, and Nora Draper. 2015. *The tradeoff fallacy: How marketers are misrepresenting American consumers an opening them up to exploitation*. Technical Report. Annenberg School for Communication, University of Pennsylvania.

19. Joseph Turow, Jennifer King, Chris Jay Hoofnagle, Amy Bleakley, and Michael Hennessy. 2009. Americans Reject Tailored Advertising and Three Activities that Enable It. *SSRN* (2009).

20. Blase Ur, Pedro Giovanni Leon, Lorrie Faith Cranor, Richard Shay, and Yang Wang. 2012. Smart, useful, scary, creepy: Perceptions of online behavioral advertising. In *Proc. SOUPS*.

21. J Warshaw, N Taft, and A Woodruff. 2016. Intuitions, Analytics, and Killing Ants: Inference Literacy of High School-educated Adults in the US. In *Proc. SOUPS*.

22. Craig E. Wills and Can Tatar. 2012. Understanding what they do with what they know. In *Proc. WPES*.

23. Craig E. Wills and Mihajlo Zeljkovic. 2011. A personalized approach to web privacy: awareness, attitudes and actions. *Information Management & Computer Security* 19, 1 (2011), 53–73.

24. Frederik J. Zuiderveen Borgesius. 2015. Improving Privacy Protection in the Area of Behavioural Targeting. *SSRN* (2015).