Web measurement for fairness and transparency

Steven Englehardt, Christian Eubank, Pete Zimmerman, Arvind Narayanan
Princeton University

Everything we do on the web is tracked, profiled, and analyzed. But what do companies do with that information? To what extent do they use it in ways that benefit us, versus unfairly discriminatory ways? While many concerns have been raised, not much is known quantitatively. We are currently building an **infrastructure to detect, measure and reverse engineer differential treatment of web users**.

Let's consider some examples. The "filter bubble" arises when algorithmic systems, such as Google search or the Facebook news feed, decide what information to show a user based on her past pattern of searches and clicks. The worry is that users will be fed reinforcing viewpoints and eventually be isolated in their own bubble. At the level of demographics, the seemingly fair principle of treating "similar" users similarly can lead to a deepening of existing disparities. Online ads have been shown to display racial bias, and online prices and deals have been shown to vary based on users' personal attributes.

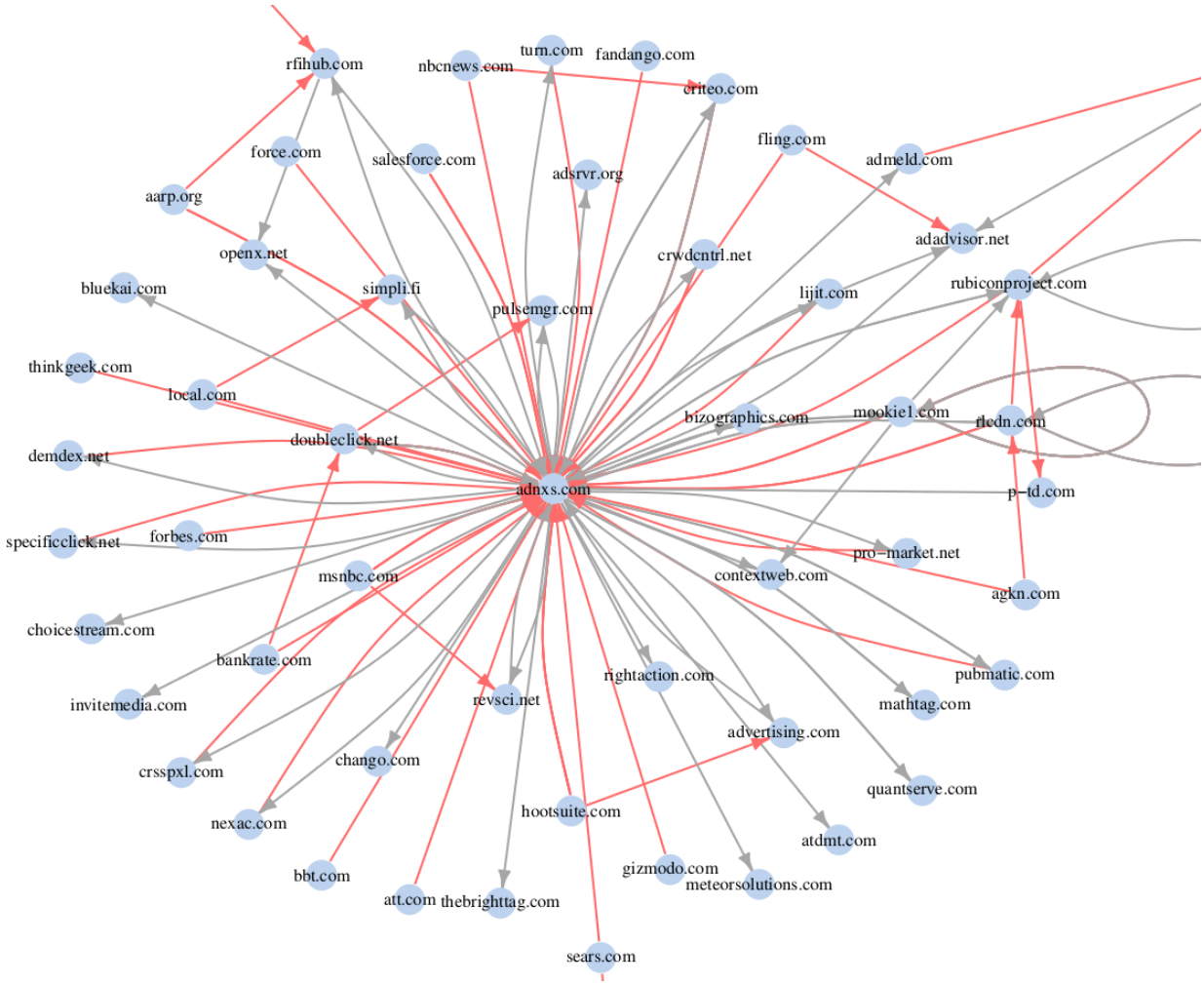
What all these and many more examples have in common is that they are ways of **using personal information for differential or discriminatory treatment**. In other words, there is a machine learning system that takes personal information as input and produces a decision as output (such as one search result versus another, or a higher price versus a lower price).

Some researchers have used manual or crowdsourcing techniques to look for such differences. While that's a great start, our approach to reverse engineering emphasizes automation, scalability, generality and speed. To this end, we're building **autonomous agents, i.e., bots, that mimic real users**. Bots with different "personas" (that vary on age, gender, affluence, location, interests, and many other attributes) browse the web, carry out searches, and so forth over a period of time. As they do so, they compare the search results, prices, ads, offers, emails, and other content they receive. A *single extensible infrastructure* with various plugins allows measuring different types of personalization or unfair discrimination across different sites.

The measurement platform draws heavily from diverse areas of computer science. We are using machine learning for building profiles of simulated users based on real user logs. Interpreting what we're seeing behind the scenes requires developing automated reverse-engineering techniques that are elaborated below. Finally, our long-term goal is to be able to run the tool on a web scale to publish a frequently-updated "census" of online privacy and discrimination. Successfully deploying such a platform is a significant systems research challenge. With this in mind, we have made our design highly modular so that different researchers can work on different parts of the infrastructure.

One particular sub-goal that we've spent much of our efforts on is **automated reverse engineering**. There is encoded information about users that's stored and transmitted via cookies and other mechanisms. Can we automatically "deobfuscate" this traffic to associate human-understandable semantics with it? For example, can we tell which values correspond to user IDs, interest segments, and other behavioral information? We are collaborating with researchers at KU Leuven on this project.

As a simple illustration of our techniques, the graph below shows a map of domains that synchronize cookies with advertising company AppNexus.[1] Cookie synchronization is a protocol by which two different third-party trackers are able to match their respective pseudonymous IDs of the user to each other, amplifying the privacy-infringing effect of online tracking.



[1] Specifically, the graph was constructed as follows. Cookie synchronization typically involves a first-party domain A embedding a third-party tracker B which redirects to another third-party tracker C. When we observe an instance of this in our web crawl data, we create a red edge from A to B and a grey edge from B to C.

Several points of note: first, this analysis is significantly deeper than tools like lightbeam for Firefox, which only observes relationships between pairs of servers. Lightbeam cannot figure out the meaning of the data that is exchanged. On the other hand, we automate the detection of cookie synchronization — this is much harder and produces much more useful results. Second, we are working on the ability to infer even more nuanced attributes such as behavioral segments and parameters related to ad auctions. Third, we are doing this measurement on a web-scale rather than a personal tool for a single user. Our goal is a **web privacy census** which will be a comprehensive map of which entities are collecting what information, what they are inferring from it, and who they are sharing it with. It is an important step in our ultimate goal of figuring out how users are treated based on that information.

It is our hope that bringing transparency to the currently invisible collection and use of personal data online will lead to greater public awareness and a more informed debate on the merits and dangers of these practices. In the case of particularly inappropriate uses of personal data, our measurement infrastructure could aid regulatory action. At present, online trackers operate at an unacceptable level of obscurity. We view our transparency initiative as a key component of digital democracy.