New tool makes online personal data more transparent

August 18 2014

The web can be an opaque black box: it leverages our personal information without our knowledge or control. When, for instance, a user sees an ad about depression online, she may not realize that she is seeing it because she recently sent an email about being sad. Roxana Geambasu and Augustin Chaintreau, both assistant professors of computer science at Columbia Engineering, are seeking to change that, and in doing so bring more transparency to the web. Along with their PhD student, Mathias Lecuyer, the researchers have developed XRay, a new tool that reveals which data in a web account, such as emails, searches, or viewed products, are being used to target which outputs, such as ads, recommended products, or prices. They will be presenting the prototype, which is designed to make the online use of personal data more transparent, at USENIX Security on August 20. The researchers have posted the open source system, as well as their findings, online for other researchers interested in studying how web services use personal data to leverage and extend.

"Today we have a problem: the web is not transparent. We see XRay as an important first step in exposing how websites are using your personal data," says Geambasu, who is also a member of Columbia's Institute for Data Sciences and Engineering's Cybersecurity Center.

We live in a "big data" world, where staggering amounts of personal data—our locations, search histories, emails, posts, photos, and more—are constantly being collected and analyzed by Google, Amazon, Facebook, and many other web services. While harnessing big data can

certainly improve our daily lives (Amazon offerings, Netflix suggestions, emergency response Tweets, etc.), these beneficial uses have also generated a big data frenzy, with web services aggressively pursuing new ways to acquire and commercialize the information.

"It's critical, now more than ever, to reconcile our privacy needs with the exponential progress in leveraging this big data," says Chaintreau, a member of the Institute for Data Sciences and Engineering's New Media Center. Geambasu adds, "If we leave it unchecked, big data's exciting potential could become a breeding ground for data abuses, privacy vulnerabilities, and unfair or deceptive business practices."

Determined to provide checks and balances on data abuse, XRay is designed to be the first fine-grained, scalable personal data tracking system for the web. For example, one can use the XRay prototype to study why a user might be shown a specific ad in Gmail. Geambasu and Chaintreau found, for example, that a Gmail user who sees ads about various forms of spiritualism might have received them because he or she sent an email message about depression.

Developing XRay was challenging, say the researchers. "The science of understanding the use of personal web data at a fine grain—looking at individual emails, photos, posts, etc.—is largely non-existent," Geambasu notes. "There really isn't anything out there that can accurately pinpoint which specific input—which search query, visited site, or viewed product—or combination of inputs explains which output. It was clear that we needed to come up with a new, robust auditing tool, one that can be applied effectively to many different services."

How it Works

"We knew from the start that our biggest challenge in achieving

transparency would be scale—how do we continue to track more data while using minimum resources?" Chaintreau says. "The theoretical results were encouraging, but seemed too good to be true. So we tested XRay in actual situations, learning from experiments we ran on Gmail, Amazon, and YouTube, and refining the design multiple times. The final design surprised us: XRay succeeded in all the experiments we ran, and it matched our theoretical predictions in increasingly complex cases. That is when we finally thought that achieving web transparency at large is not a dream in a distant future but something we can start building toward now."

The current XRay system works with Gmail, Amazon, and YouTube. However, XRay's core functions are service-agnostic and easy to instantiate for new services, and they can track data within and across services. The key idea in XRay is to use black-box correlation of data inputs and outputs to detect data use.

To assess XRay's practical value, the researchers created an XRay-based demo service that continuously collects and diagnoses Gmail ads related to a set of topics, including various diseases, pregnancy, race, sexual orientation, divorce, debt, etc. They created emails that included keywords closely related to one topic and then launched XRay's Gmail ad collection and examined the targeting associations. XRay's data is now available online to anyone interested in sensitive-topic ad targeting in Gmail.

"We've just started to peek into XRay's targeting data and even at this early stage, we've seen a lot of interesting behaviors," Geambasu says. "We know that we need larger-scale experience to formalize and quantify our conclusions, but we can already make several interesting observations."

The researchers note that (1) It is definitely possible to target sensitive

topics in users' inboxes, including cancer, depression, or pregnancy. (2) For many ads, targeting was extremely obscure and non-obvious to end-users, which opens them up to abuses. (3) The researchers have already seen signs of such abuses, for instance, a number of subprime loan ads for used cars targeting debt in users' inboxes. Examples of ads and their targeted topics can be found on the XRay website.

The tool can be used to increase user awareness about how their data is being used, as well as provide much needed tools for auditors, such as researchers, journalists, and investigators, to keep that use under scrutiny. Geambasu and Chaintreau, who recently won a Magic Grant from the Brown institute for Media Innovation to build better transparency tools, have made the XRay prototype available for auditors at http://xray.cs.columbia.edu.

"Our work calls for and promotes the best practice of voluntary transparency," says Chaintreau, "while at the same time empowering investigators and watchdogs with a significant new tool for increased vigilance, something we need more of every day."

Provided by Columbia University