New tool expands tracking of personal data on the Web

October 9 2015

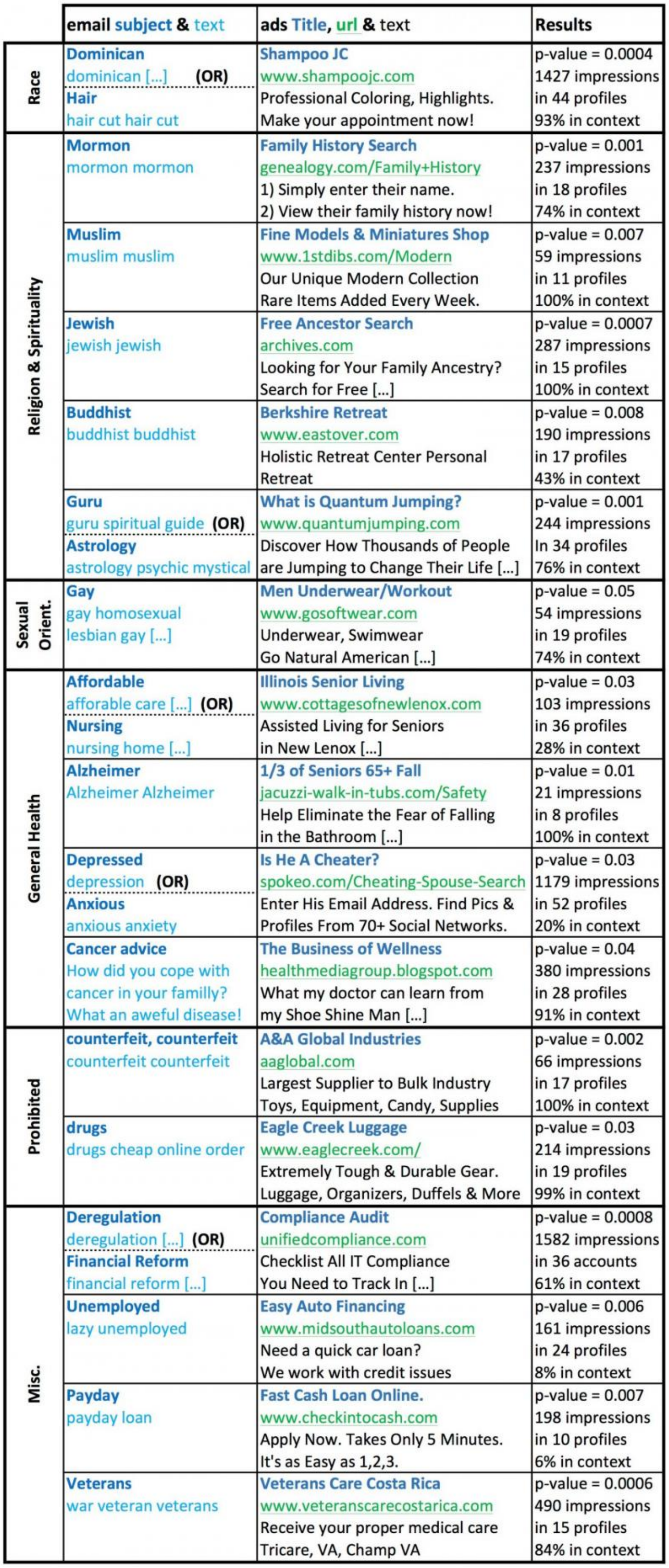
A second-generation tool called Sunlight is intended to bring greater transparency to the Web. Credit: (Tumblr)

Navigating the Web gets easier by the day as corporate monitoring of our emails and browsing habits fine-tune the algorithms that serve us personalized ads and recommendations. But convenience comes at a cost. In the wrong hands, our personal information can be used against us, to discriminate on housing and health insurance, and overcharge on goods and services, among other risks.

"The Web is like the Wild West," says Roxana Geambasu, a computer scientist at Columbia Engineering and the Data Science Institute. "There's no oversight of how our data are being collected, exchanged and used."

With computer scientists, Augustin Chaintreau and Daniel Hsu, and graduate students Mathias Lecuyer, Riley Spahn and Yannis Spiliopoulos, Geambasu has designed a second-generation tool for bringing transparency to the Web. It's called Sunlight and builds on its predecessor, XRay, which linked ads shown to Gmail users with text in their emails, and recommendations on Amazon and YouTube with their shopping and viewing patterns. The researchers will present the new tool and a related study on Oct. 14 in Denver, at the Association for Computing Machinery's annual conference on security.

Sunlight works at a wider scale than XRay, and more accurately matches user-tailored ads and recommendations to tidbits of information supplied by users, the researchers say. Prior researchers have traced specific ads, product recommendations and prices to specific inputs like location, search terms and gender, one by one. One tool, AdFisher, received attention earlier this year after showing that fake Web users thought to be male job seekers were more likely than female job seekers to be shown ads for executive jobs when later visiting a news site.



	email subject & text	ads Title, url & text	Results
Race	Dominican dominican […] (OR) Hair hair cut hair cut	Shampoo JC www.shampoojc.com Professional Coloring, Highlights. Make your appointment now!	p-value = 0.0004 1427 impressions in 44 profiles 93% in context
Religion & Spirituality	Mormon mormon mormon	Family History Search genealogy.com/Family+History 1) Simply enter their name. 2) View their family history now!	p-value = 0.001 237 impressions in 18 profiles 74% in context
	Muslim muslim muslim	Fine Models & Miniatures Shop www.1stdibs.com/Modern Our Unique Modern Collection Rare Items Added Every Week.	p-value = 0.007 59 impressions in 11 profiles 100% in context
	Jewish jewish jewish	Free Ancestor Search archives.com Looking for Your Family Ancestry? Search for Free […]	p-value = 0.0007 287 impressions in 15 profiles 100% in context
	Buddhist buddhist buddhist	Berkshire Retreat www.eastover.com Holistic Retreat Center Personal Retreat	p-value = 0.008 190 impressions in 17 profiles 43% in context
	Guru guru spiritual guide (OR) Astrology astrology psychic mystical	What is Quantum Jumping? www.quantumjumping.com Discover How Thousands of People are Jumping to Change Their Life […]	p-value = 0.001 244 impressions In 34 profiles 76% in context
Sexual Orient.	Gay gay homosexual lesbian gay […]	Men Underwear/Workout www.gosoftwear.com Underwear, Swimwear Go Natural American […]	p-value = 0.05 54 impressions in 19 profiles 74% in context
General Health	Affordable afforable care […] (OR) Nursing nursing home […]	Illinois Senior Living www.cottagesofnewlenox.com Assisted Living for Seniors in New Lenox […]	p-value = 0.03 103 impressions in 36 profiles 28% in context
	Alzheimer Alzheimer Alzheimer	1/3 of Seniors 65+ Fall jacuzzi-walk-in-tubs.com/Safety Help Eliminate the Fear of Falling in the Bathroom […]	p-value = 0.01 21 impressions in 8 profiles 100% in context
	Depressed depression (OR) Anxious anxious anxiety	Is He A Cheater? spokeo.com/Cheating-Spouse-Search Enter His Email Address. Find Pics & Profiles From 70+ Social Networks.	p-value = 0.03 1179 impressions in 52 profiles 20% in context
	Cancer advice How did you cope with cancer in your familly? What an aweful disease!	The Business of Wellness healthmediagroup.blogspot.com What my doctor can learn from my Shoe Shine Man […]	p-value = 0.04 380 impressions in 28 profiles 91% in context
Prohibited	counterfeit, counterfeit counterfeit counterfeit	A&A Global Industries aaglobal.com Largest Supplier to Bulk Industry Toys, Equipment, Candy, Supplies	p-value = 0.002 66 impressions in 17 profiles 100% in context
	drugs drugs cheap online order	Eagle Creek Luggage www.eaglecreek.com/ Extremely Tough & Durable Gear. Luggage, Organizers, Duffels & More	p-value = 0.03 214 impressions in 19 profiles 99% in context
Misc.	Deregulation deregulation […] (OR) Financial Reform financial reform […]	Compliance Audit unifiedcompliance.com Checklist All IT Compliance You Need to Track In […]	p-value = 0.0008 1582 impressions in 36 accounts 61% in context
	Unemployed lazy unemployed	Easy Auto Financing www.midsouthautoloans.com Need a quick car loan? We work with credit issues	p-value = 0.006 161 impressions in 24 profiles 8% in context
	Payday payday loan	Fast Cash Loan Online. www.checkintocash.com Apply Now. Takes Only 5 Minutes. It's as Easy as 1,2,3.	p-value = 0.007 198 impressions in 10 profiles 6% in context
	Veterans war veteran veterans	Veterans Care Costa Rica www.veteranscarecostarica.com Receive your proper medical care Tricare, VA, Champ VA	p-value = 0.0006 490 impressions in 15 profiles 84% in context

Columbia researchers found evidence that some Gmail ads appeared to contradict Google's ban on using sensitive information in targeted ads. Credit: Columbia University

Sunlight, by contrast, is the first to analyze numerous inputs and outputs together to form hypotheses that are tested on a separate dataset carved out from the original. At the end, each hypothesis, and its linked input and output, is rated for statistical confidence. "We're trying to strike a balance between statistical confidence and scale so that we can start to see what's happening across the Web as a whole," said Hsu.

The researchers set up 119 Gmail accounts, and over a month last fall sent 300 messages with sensitive words in the subject line and body of the email. About 15 percent of the ads that followed appeared to be targeted; some seemed to contradict Google's policy to not target ads based "on race, religion, sexual orientation, health or sensitive financial categories," the researchers said. For example, words typed into the subject line of a message— "unemployed," "depressed," and "Jewish," were found to trigger ads for "easy auto financing," a service to find "cheating spouses," and a "free ancestor" search, respectively.

The researchers also set up fake browsing profiles and surfed the 40 most popular sites on the Web to see what ads popped up. They found that just 5 percent of the ads appeared to be targeted, but some seemed to violate Google's advertising ban on products and services facilitating drug use, they said. For example, a visit to "hightimes.com" triggered an ad for bongs at AquaLab Technologies, researchers said. Interestingly, the algorithms also seemed to pick up on the political leanings of popular news sites, pitching Israeli bonds to Fox News readers, and an anti-Tea

Party candidate to Huffington Post readers.

The researchers caution against inferring that Google and other companies are intentionally using sensitive information to target ads and recommendations. The flow of personal data on the Web has become so complex, they said, that companies themselves may not know how targeting is taking place.

In Nov. 10, 2014, Google abruptly shut down Gmail ads - the last day that Geambasu and her colleagues were able to collect data. The ads appear to have been replaced by so-called organic ads displayed in the promotions tab. Sunlight has the ability to detect targeting in those ads, too, said Geambasu, but the researchers haven't yet given that a try.

Sunlight's intended audience is regulators, consumer watchdogs and journalists. The tool lets them explore how personal information is being used and decide where closer investigation is needed, they said. "In many ways the Web has been a force for good, but there needs to be accountability if it's going to remain that way," said Chaintreau.

"Sunlight is distinctive in that it can examine multiple types of inputs simultaneously (e.g., gender, age, browsing activity) to develop hypotheses about which of these inputs impact certain outputs (e.g., ads on Gmail)," said Anupam Datta, a researcher at Carnegie Mellon who led the development of the AdFisher tool and was not involved in the current study. "This tool takes us closer to the critical goal of discovering personal data use effects at scale."

More information: www.cs.columbia.edu/~djhsu/papers/sunlight.pdf

Provided by Columbia University

Citation: New tool expands tracking of personal data on the Web (2015, October 9) retrieved 3 September 2024 from https://phys.org/news/2015-10-tool-tracking-personal-web.html