Tracking the Trackers:
A Large-Scale Analysis of Embedded Web Trackers

Sebastian Schelter
Technische Universität Berlin
sebastian.schelter@tu-berlin.de

Jérôme Kunegis
Universität Koblenz-Landau
kunegis@uni-koblenz.de

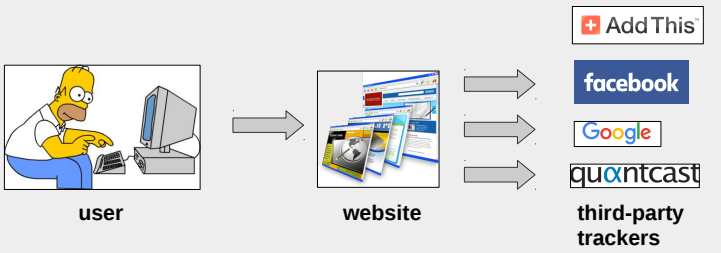
“You never browse alone”: 45% of all Websites contain Trackers

How does Online Tracking work?

- website you visit embeds resources from third-party trackers in the HTML code
- your browser loads the resources from the servers of the tracking company
- trackers can record the site you visit, your IP, referrer URI, operating system, compute a browser fingerprint, and set cookies

Privacy Hazards

- users recognized across many websites
- news consumption is recorded
- tracking on health-related sites
- intelligence agencies piggyback on tracking infrastructure to build databases of surfing behavior of millions of people



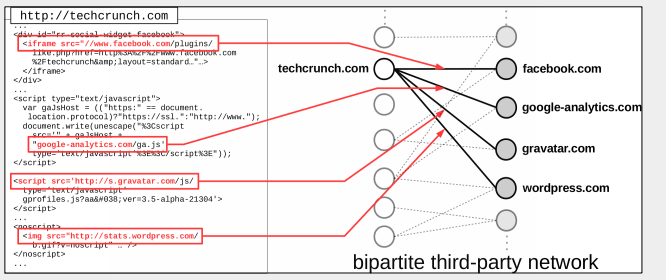
Our Dataset

Web Trackers in CommonCrawl 2012

- CommonCrawl 2012 is a publicly available web crawl consisting of more than **3.5 billion webpages**
- developed an extractor that finds third-party resources in HTML code, looking at *iframe*, *script* and *image* tags, as well as *JavaScript* variables

Parsing 200+ Terabytes of Web Data

- created a parallel Hadoop implementation of our extractor
- ran it on the whole CommonCrawl 2012 corpus using the Amazon Cloud
- extracted domains of websites and trackers to form the ***bipartite tracking graph***



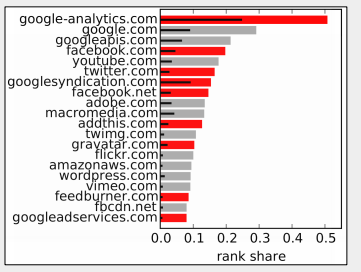
Web Tracking as a Global Phenomenon

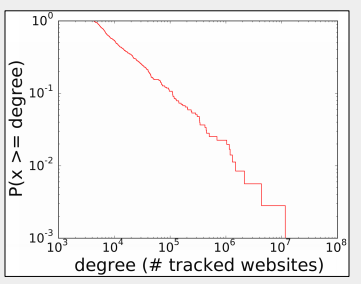
Ranking Third-Parties

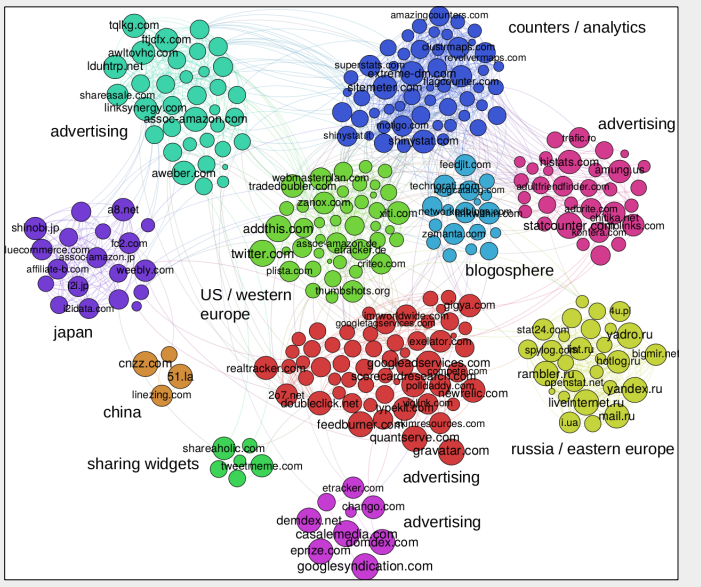
- **rank share** of third-party *t* in domain set *D:* sum of the PageRanks of domains from *D* that have *t* embedded, normalized by the sum of the pageranks of domains in *D*
- **9 out of the 20 top third-parties are tracking services**

Distributional Patterns

- distribution of the number of sites tracked per tracker follows power-law
- clustering of the tracker co-occurrence graph by modularity maximization reveals country-specific and category-specific patterns







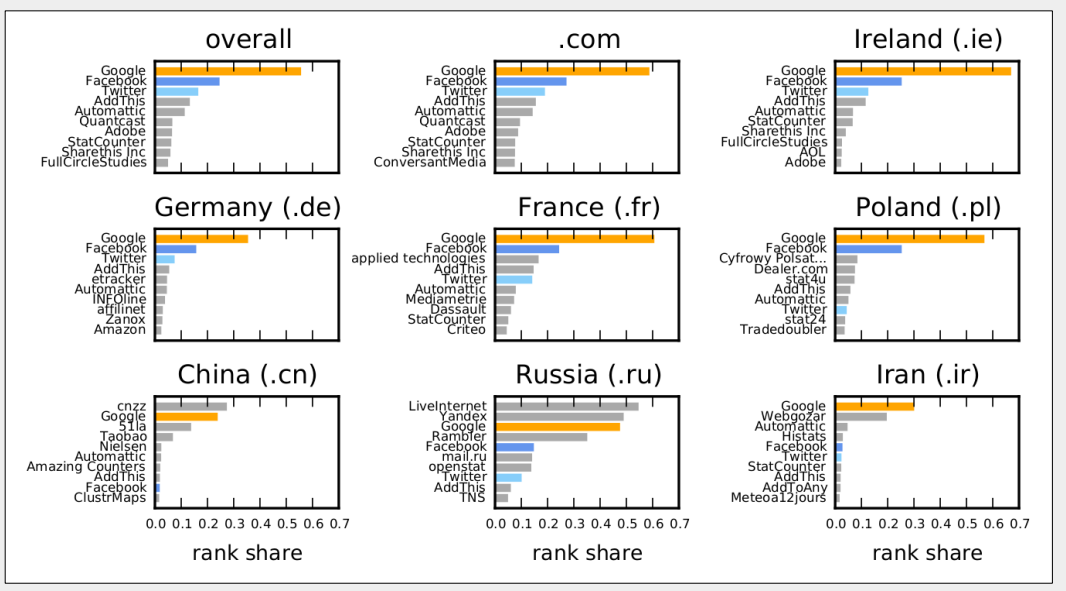
Web Tracking as a Local Phenomenon

Top Trackers per Country

- computed top tracking companies per country using TLD as proxy
- **small set of US-based companies dominate tracking globally**
- typically accompanied by a couple of domestic tracking companies
- exceptions: China, Russia, Iran, Ukraine

Correlation Analysis of Tracking Domination

- computed correlation of tracking dominance with various political, socio-cultural and economic factors
- found strong and significant **correlation with political factors such as *freedom of the press***, no significant correlation with economic factors such as *online ad spending per capita*



dataset available at **http://konect.uni-koblenz.de/networks/trackers-trackers**