Preserving Needles in the Haystack

A search engine and forensic documentation system for privacy violations on the web

Timothy Libert, CyLab Partners Conference, 2020 | CONFIDENTIAL, NOT FOR DISTRIBUTION

Background: Web Tracking

- When you visit websites “third-parties” are able to link your browsing behavior to individual advertising profiles by using IP addresses, cookies, and other methods.
- Web browsing can reveal sensitive personal information (eg mental health diagnosis), decades of surveys have shown the public does not approve of it, and decades of technical research show you can’t avoid it even if you want to.
- One of the biggest problems in the world today is citizens losing trust in *institutions*, so this isn’t just about ads, or even privacy, it’s about people *no longer viewing technology as something that improves their lives*.
- This is a huge problem, and as technologists, it’s our problem to fix.

How to stop web tracking?

- We are stuck in a 20+ year stalemate by which privacy researchers invent new tracker blockers and much better funded adtech companies come out with new trackers and tracker-blocker-*blockers* (rinse/repeat).
- New laws, such as GDPR and CCPA, are designed to reign in bad behavior online with strong fines, as the best chance of stopping tracking is making it unprofitable.
- As of yet GDPR has been more smoke than fire, and while the *law* is strong *enforcement* is weak. This may be partially attributed to the complexity of cases and limited budgets.
- ***My question: How can we use automation to <u>create superhuman regulators</u> rather than short-lived blocking tools?***

Supercharging Regulation and Litigation

- Two main ways automation can help in regulation and litigation: 1) selecting targets; 2) admitting evidence to court.
- Selecting a target can be a matter of finding a very specific type of violation at the intersection of technology and policy, and then picking the best case among many to set a precedent or win a settlement.
- Admitting evidence to court comes down to forensics, can you show the court proof that a privacy violation did, in fact, happen? Preferably in a “clean room” environment?

The Web Privacy Haystack: “Collect it All”

“Let’s collect the whole haystack. Collect it all, tag it, store it. . . . And whatever it is you want, you go searching for it.”

Former NSA Director Keith Alexander, Quoted in Washington Post

Privacy Violation Search Engine

- Ever since I read the Alexander quote I wanted to become the NSA of web tracking.
- For a privacy violation search engine we need to collect as much data on tracking as we can. In other words, the entire haystack.
- This requires two things:
  - Browser automation for measuring privacy violations
  - Scaling and speed

Browser Automation

- For browser automation we leverage Chrome to record and store:
    - network events
    - cookies
    - SSL certificates
    - screen shots
    - page text
    - Javascript files
    - XHR responses
    - …and more!
- Everything is timestamped with millisecond resolution and files are hashed so they can be admitted as evidence in court.

Scaling and Speed

- Goal is to scan as many sites as possible which means going really fast.
- This means we need to optimize how quickly our browsers can extract tracking data, and distribute the task across many computers (some potentially in other countries).
- To speed up search queries we perform extensive preprocessing at ingest time so data is easily accessible.
- We have developed a novel distributed system that currently runs several hundred instances of Chrome in parallel on the CMU campus as well as some remote nodes in the EU and China.

Search Space

- Performed main scan this summer, took roughly 5 weeks with < $20k of hardware
- 11M stateful page loads conducted across 2.3M sites
- 1.4B requests
- 300M cookies
- 1M policies (privacy, terms of service, etc)
- Nearly 2B words of policy text, roughly 15 years to read w/no breaks
- Much more!

Preserving a Needle

The Needles

- Finding our needles has two components, the technical observation (eg a cookie) and the policy component which can either be a law (eg COPAA, GDPR) or a legal promise made by the site (eg Terms of Service, Privacy Policy).
- We can search for pages covering a given topic, making certain types of claims, or have trackers from specific companies. Once we find them we can document the behavior, archive it, and reverse engineer it.
- Let’s do an example…

Health Privacy

- Health is a universally recognized privacy-sensitive context and is highly regulated.
- In the US there is also tons of money to be made in micro-targeting pharmaceutical ads by leveraging web browsing data.
- Perhaps a regulator or class-action litigant should be paying attention to this?
- Let’s find some targets, create a forensic record, and reverse engineer this!

Health Privacy

- Let’s start to dig:
  - Sites that have a policy mentioning “HIPAA”: 9,077
  - Sites mentioning “HIPAA” in a policy that specifically cover mental health: 207
- Interesting start, but let’s find something even more specific.

Mental Health Ad Targeting

- “Company X” has a policy that states it “does not knowingly create [advertising] segments that are based upon…sensitive health information”
- Yet deep within their site there is a PDF file that reveals they allow targeting on mental health conditions
- **This seems like a deceptive legal claim, let’s find enforcement targets!**

Mental Health Ad Targeting

- Search Results:
  - Mental health websites making a third-party request to Company X: 24
  - Mental health websites mentioning “HIPAA” in a policy making a third-party request to Company X: 3
- A regulator or litigator has 24 potential targets for an action against Company X for making deceptive claims about health privacy, 3 of which have the *added bonus* of a deceptive HIPAA claim by the site.

Forensics: Capturing Company X

- Site “S” is very highly ranked website for mental health has trackers from Company X.
- There is some uncertainty as to how HIPAA may apply to the web in the United States on a generalized basis, but there is *no uncertainty* health is protected by GDPR or that setting cookies requires affirmative opt-in consent.
- Our orchestration API told a scan node in the EU to do a forensic capture of a page about anxiety tracked by Company X on Site S.
- This is essentially a “clean room” measurement from a residential IP, in the EU, and thus court-admissible.
- Reverse engineering the forensic data shows three companies are involved in this tracking via a series of redirects and iFrames, expanding the actionable scope for litigation and enforcement

Summary

- Let’s stop trying to *block trackers* and use technology to *create superhuman regulators*.
- My system allows for large scale searching and fine-grained documentation of privacy violations on the web.
- Much of this work supported by CyLab seed funding, thanks!
- Q/A?