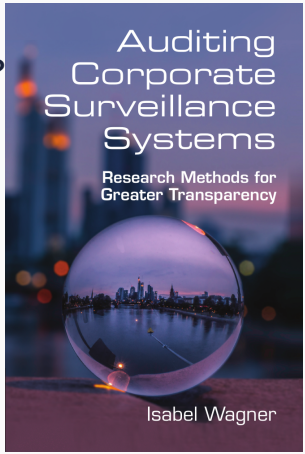
HOW CAN WE STUDY CORPORATE SURVEILLANCE?

OVERVIEW OF METHODS AND RESULTS FOR THE EXAMPLE OF STATEFUL TRACKING

Isabel Wagner

De Montfort University



OVERVIEW

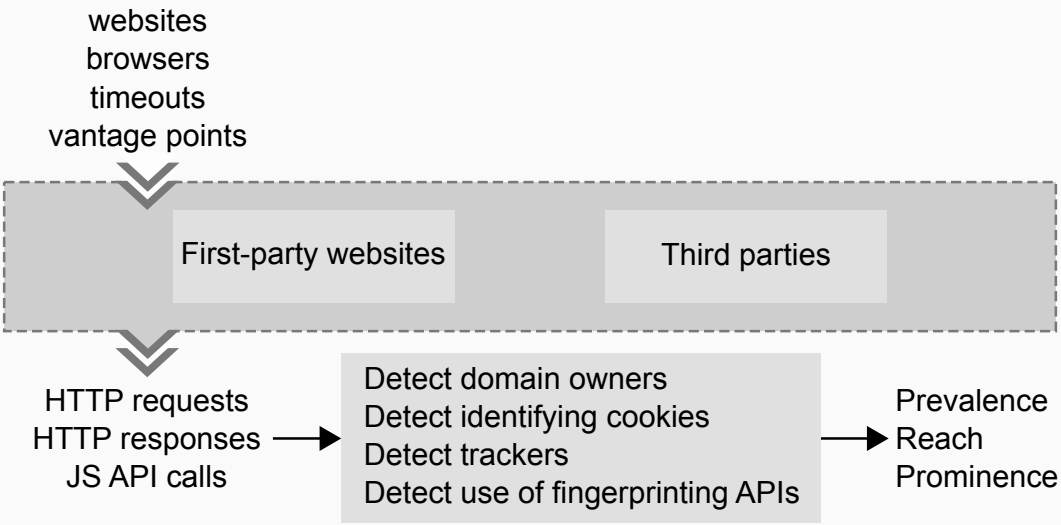
- Research questions: what specifically do we want to find out about stateful tracking?
- Experiment design: how can we set up experiments to answer our questions?
- Response variables: which indicators can we measure in our experiments?
- Results: what have researchers already found out about stateful tracking?

Study designs

RESEARCH QUESTIONS

- How does tracking work? Which tracking methods are there?
- How common is tracking?
- How many websites are the biggest trackers embedded in?
- Has tracking changed after new data protection regulations came into effect, and if so, how?

DESIGN FOR STUDYING STATEFUL TRACKING



Heuristics

WHICH THIRD PARTIES ARE TRACKERS?

- Heuristics to classify which cookies carry user identifiers[1]
  - Length of cookie value: 8-100 characters
  - Expiration date: 90 days or more
  - Value for each browser constant, but different between browsers, e.g. quantified via Ratcliff/Obershelp similarity (number of matching characters divided by the total number of characters in the two strings)
- Detect which third parties are trackers[2]
  - Rely on lists curated by ad-/tracker-blockers, e.g. EasyList, EasyPrivacy
- Detect which third party domains belong to the same corporation
  - Whois lookup, manual website lookup, TLS certificates, chrunchbase, hoovers

[1] G. Acar, C. Eubank, S. Englehardt, *et al.*, "The Web Never Forgets: Persistent Tracking Mechanisms in the Wild," in *Proceedings of the 2014 ACM SIGSAC Conference on Computer and Communications Security*, ser. CCS '14, Scottsdale, Arizona, USA: ACM, 2014, pp. 674–689. DOI: *10.1145/2660267.2660347*.

[2] S. Englehardt and A. Narayanan, "Online Tracking: A 1-million-site Measurement and Analysis," in *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security*, ser. CCS '16, Vienna, Austria: ACM, 2016, pp. 1388–1401. DOI: *10.1145/2976749.2978313*.

Quantitative measures

MEASURES FOR TRACKING

- Prevalence of tracker use, e.g. number or percentage of third-party requests per app/website[3]
- Reach of tracking companies, e.g. percentage of websites tracked by a specific tracker[4]
- Prominence of tracking companies: similar to reach, but de-emphasizes trackers on low-ranked websites[5]
- Penetration of trackers, e.g. percentage of users who send at least one request to a specific tracker within n days[6]

[3] G. Merzdovnik, M. Huber, D. Buhov, *et al.*, "Block Me If You Can: A Large-Scale Study of Tracker-Blocking Tools," in *2017 IEEE European Symposium on Security and Privacy (EuroS P)*, Paris, France: IEEE, Apr. 2017, pp. 319–333. DOI: *10.1109/EuroSP.2017.26*.

[4] T. Libert, "An Automated Approach to Auditing Disclosure of Third-Party Data Collection in Website Privacy Policies," in *Proceedings of the 2018 World Wide Web Conference*, ser. WWW '18, Lyon, France: International World Wide Web Conferences Steering Committee, 2018, pp. 207–216. DOI: *10.1145/3178876.3186087*.

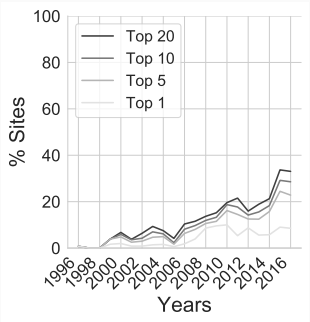
[5] S. Englehardt and A. Narayanan, "Online Tracking: A 1-million-site Measurement and Analysis," in *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security*, ser. CCS '16, Vienna, Austria: ACM, 2016, pp. 1388–1401. DOI: *10.1145/2976749.2978313*.

[6] H. Metwalley, S. Traverso, M. Mellia, *et al.*, "The Online Tracking Horde: A View from Passive Measurements," in *Traffic Monitoring and Analysis*, M. Steiner, P. Barlet-Ros, and O. Bonaventure, Eds., ser. Lecture Notes in Computer Science, Cham: Springer, 2015, pp. 111–125.

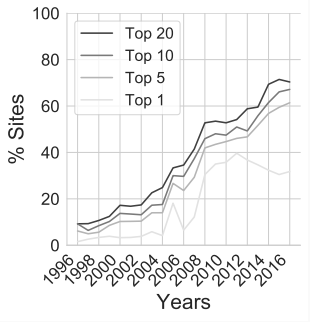
Results

HISTORIC USE OF TRACKING

- Use Internet Archive's Wayback Machine to study use of tracking APIs on 500 most popular sites since 1996[7]
- Sites that contact at least 5 trackers: 5% in 2000, 40% in 2016
- Top 5 trackers reach <10% of sites in 2000, 60% in 2016



Trackers



Third parties

[7] A. Lerner, A. K. Simpson, T. Kohno, *et al.*, "Internet Jones and the Raiders of the Lost Trackers: An Archaeological Study of Web Tracking from 1996 to 2016," in *25th USENIX Security Symposium (USENIX Security 16)*, Austin, TX, USA: USENIX, 2016

PREVALENCE OF TRACKING

- Top 10k websites (2015)[8]
  - Top 5 trackers reach 30% of websites
- Top 1m websites (2016)[9]
  - Top 5 third-party domains all owned by Google/Alphabet
  - Top tracker reaches >60% of sites

[8] T.-C. Li, H. Hang, M. Faloutsos, *et al.*, "TrackAdvisor: Taking Back Browsing Privacy from Third-Party Trackers," in *Passive and Active Measurement*, J. Mirkovic and Y. Liu, Eds., ser. Lecture Notes in Computer Science, Cham: Springer, 2015, pp. 277–289.

[9] S. Englehardt and A. Narayanan, "Online Tracking: A 1-million-site Measurement and Analysis," in *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security*, ser. CCS '16, Vienna, Austria: ACM, 2016, pp. 1388–1401. DOI: *10.1145/2976749.2978313*.

PREVALENCE OF COOKIE SYNCHRONIZATION

- 5k websites visited by human users in 2016[10]
  - 96% of user-identifying cookies are shared between 2 or more third parties
- Top 100k websites[11]
  - Doubleclick is most prolific cookie sync'er
  - 90% of top-50 third parties and 85% of top-100 share cookies with at least one other third party
- 1-year traffic from 850 human users[12]
  - 20% of users experience first cookie sync within first day, 38% in first week
  - 1 sync per 68 HTTP GET requests
  - Each ID cookie is shared with median of 3.5 third parties

[10] M. Falahrastegar, H. Haddadi, S. Uhlig, *et al.*, "Tracking Personal Identifiers Across the Web," in *Passive and Active Measurement*, T. Karagiannis and X. Dimitropoulos, Eds., ser. Lecture Notes in Computer Science, Cham: Springer, 2016, pp. 30–41.

[11] S. Englehardt and A. Narayanan, "Online Tracking: A 1-million-site Measurement and Analysis," in *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security*, ser. CCS '16, Vienna, Austria: ACM, 2016, pp. 1388–1401. DOI: *10.1145/2976749.2978313*.

[12] P. Papadopoulos, N. Kourtellis, and E. Markatos, "Cookie Synchronization: Everything You Always Wanted to Know But Were Afraid to Ask," in *The World Wide Web Conference*, ser. WWW '19, San Francisco, CA, USA: ACM, 2019, pp. 1432–1442. DOI: *10.1145/3308558.3313542*.

PREVALENCE OF CROSS-DEVICE TRACKING

- Study with real traffic from 126 users[13]
  - Detect 2.5k unique third-party trackers
  - 124 cross-device trackers from 87 companies
  - Trackers classified as CDT if tracker occurs on mobile and desktop sites and company website confirms that they perform CDT
- Study with synthetic traffic: 100 websites on two devices[14]
  - 861 third-party domains receive data that is useful for probabilistic CDT
  - 106 third-party domains share cookie identifiers that enable sharing of CDT graphs between third parties
  - 16% of sites share PII which can be used for deterministic CDT

[13] S. Zimmeck, J. S. Li, H. Kim, *et al.*, "A Privacy Analysis of Cross-device Tracking," in *26th USENIX Security Symposium*, Vancouver, BC, Canada: USENIX, Aug. 2017, p. 19.

[14] J. Brookman, P. Rouge, A. Alva, *et al.*, "Cross-Device Tracking: Measurement and Disclosures," *Proceedings on Privacy Enhancing Technologies*, vol. 2017, no. 2, pp. 133–148, Apr. 2017. DOI: *10.1515/popets-2017-0020*.

TRACKING VIA TLS SESSION RESUMPTION

- TLS session tickets allow to resume prior TLS sessions
  - Avoid costly TLS handshake (especially on server side)
  - Server can send “lifetime hint” for sessions, up to 7 days (TLS 1.3)
  - First parties can track users based on TLS session IDs
  - Third parties can track users across sites because third-party TLS sessions can be resumed in the context of different first-party sites
- Session lifetimes, tracking duration on the web?[15]
  - Record TLS handshakes for Alexa top 1 million, daily for 30 days
  - Revisit in 5-minute intervals to probe session resumption lifetime
  - 80% of lifetimes < 10 minutes, but big tech uses 28-48 hours
  - With 7d lifetime, 65% of users can be tracked permanently; with 24h: 1.3%

[15] E. Sy, C. Burkert, H. Federrath, *et al.*, “Tracking Users Across the Web via TLS Session Resumption,” in *Proceedings of the 34th Annual Computer Security Applications Conference*, ser. ACSAC ’18, San Juan, PR, USA: ACM, 2018, pp. 289–299. DOI: 10.1145/3274694.3274708.

Policy issues: tracking vs. regulation

MARKET CONCENTRATION OF TRACKERS

- Is the tracking market a case for competition regulators?[16]
- Herfindahl-Hirschman index: $\text{HHI} = \sum_{i=1}^{N} s_i^2$ (where s_i is the market share of tracker i)
  - How to compute tracker market share?
    - Proportion of prevalence or proportion of prominence
  - EU threshold: concentration of 0.1, consolidation with increase of 0.025
  - US threshold: concentration of 0.25
- Study with top 5k websites
  - Taking parent/subsidiary relationships into account, HHI is 0.12
  - 2007 acquisition of Doubleclick by Google: increase of HHI by 0.039

[16] R. Binns, J. Zhao, M. V. Kleek, *et al.*, "Measuring Third-party Tracker Power Across Web and Mobile," *ACM Trans. Internet Technol.*, vol. 18, no. 4, 52:1–52:22, Aug. 2018. DOI: *10.1145/3176246*.

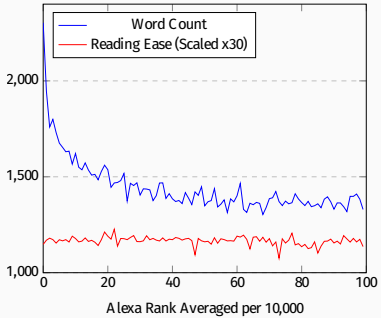
CROSS-BORDER TRACKING

- To what extent do tracking data flows cross data protection borders?[17]
  - E.g. EU border (GDPR)
  - Crowdsourcing study with 350 users
- Relies on correct geolocation of users and tracking servers
  - Collect IP addresses for tracking domains, then use RIPE IPmap
- 84% of tracking traffic originating in the EU terminates in the EU
  - 10% go to North America
- 90% of tracking traffic from South America terminates in North America
- Results hold for traffic in sensitive categories (health etc.)

[17] C. Iordanou, G. Smaragdakis, I. Poese, *et al.*, "Tracing Cross Border Web Tracking," in *Proceedings of the Internet Measurement Conference 2018*, ser. IMC '18, Boston, MA, USA: ACM, 2018, pp. 329–342. DOI: 10.1145/3278532.3278561.

DATA FLOWS VS. PRIVACY POLICIES

- How many privacy policies disclose which third parties they contact?[18]
  - Parse policies + measure traffic of top 1m websites
  - <15% of trackers are disclosed
  - 38% of sites disclose tracking by Google
- How readable are privacy policies?
  - Flesch Reading Ease score: scores > 45 are “easy to read”
  - Average score for privacy policies: 39.8
  - Score for trackers’ privacy policies: 35.4
  - Average policy length: 3,882 words (15-minute read)
  - But often refer users to third-party privacy policies



[18] T. Libert, “An Automated Approach to Auditing Disclosure of Third-Party Data Collection in Website Privacy Policies,” in *Proceedings of the 2018 World Wide Web Conference*, ser. WWW ’18, Lyon, France: International World Wide Web Conferences Steering Committee, 2018, pp. 207–216. DOI: 10.1145/3178876.3186087

PRE- VS. POST-GDPR (1)

- Prevalence of tracking on 200 European news sites[19]
  - 22% fewer third-party cookies April-July 2018
  - 7% less social media content
  - But: reach of top tracker (Google) only dropped by 1% (97% to 96%)
- Differences in cookie setting depending on user location?[20]
  - 53% of top 100k websites do not set cookies
  - 26% of cookie-setting sites set cookies for US users but not EU users
  - This diversification of websites complicates measurement studies: restrict scope to one region or multiple vantage points for measurement

[19] T. Libert, L. Graves, and R. K. Nielsen, "Changes in third-party content on European news websites after GDPR," Reuters Institute for the Study of Journalism, Tech. Rep. pubs:909043, 2018.

[20] A. Dabrowski, G. Merzdovnik, J. Ullrich, *et al.*, "Measuring Cookies and Web Privacy in a Post-GDPR World," in *Passive and Active Measurement*, D. Choffnes and M. Barcellos, Eds., ser. Lecture Notes in Computer Science, Cham: Springer International Publishing, 2019, pp. 258–270. DOI: *10.1007/978-3-030-15986-3_17*.

PRE- VS. POST-GDPR (2)

- Cookie notices[21]
  - Study 1.5k websites from 18 vantage points
  - 40% of websites display cookie notice, regardless of user location
  - Instead differentiation based on website's top-level domain: for .com domains, EU users have 32% increased odds of seeing cookie notice
- Consent notices[22]
  - Visual properties: position, type of choice offered, framing of content
  - Nudges and default selections strongly impact user choice
  - Websites employ dark patterns for obtaining consent -> neither explicit nor informed (as required by GDPR)

[21] R. van Eijk, H. Asghari, P. Winter, *et al.*, "The Impact of User Location on Cookie Notices (Inside and Outside of the European Union)," in *IEEE Security & Privacy Workshop on Technology and Consumer Protection (ConPro '19)*, San Francisco, CA, USA: IEEE, May 2019.

[22] C. Utz, M. Degeling, S. Fahl, *et al.*, "(Un)Informed Consent: Studying GDPR Consent Notices in the Field," in *Proceedings of the 2019 ACM SIGSAC Conference on Computer and Communications Security*, ser. CCS '19, London, United Kingdom: ACM, 2019, pp. 973–990. DOI: 10.1145/3319535.3354212.

Summary

STATEFUL TRACKING: METHODS AND RESULTS

- 5-step process to study corporate surveillance:
  1. Define research questions
  2. Design experiments
  3. Define input and response variables, necessary heuristics and quantitative measures
  4. Run experiments: collect data
  5. Analyze data and present results
- Results from many studies together paint picture of stateful tracking:
  - Prevalence, including historical evolution
  - Characteristics and prevalence of specific tracking techniques
  - Effects of regulations on tracking

ABOUT THIS SLIDE DECK

- These slides are designed to accompany a lecture based on the textbook “Auditing Corporate Surveillance Systems: Research Methods for Greater Transparency” by Isabel Wagner, published in 2022 by Cambridge University Press.
-
-
- Please leave this slide intact, but indicate modifications below.
  - Version 2022-04
  - Improved version for release on book website (Isabel Wagner)
- Updated versions of the original slide deck are available online: corporatesurveillance.org