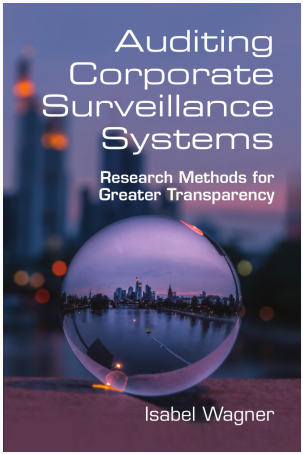
EXPERIMENT DESIGN

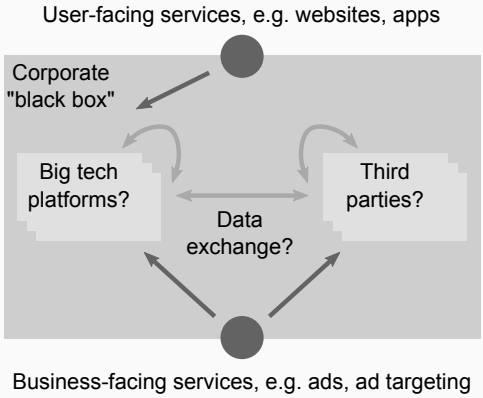
Isabel Wagner

De Montfort University



WHAT IS THE AIM OF TRANSPARENCY RESEARCH?

- Gather information about corporate surveillance from the outside
  - Methods for studying black-box systems
  - Real-world implications for system design and regulation
- Typically experimental:
  - Systematically interact with the services corporations offer to users and/or businesses
  - To infer how these services work on the inside



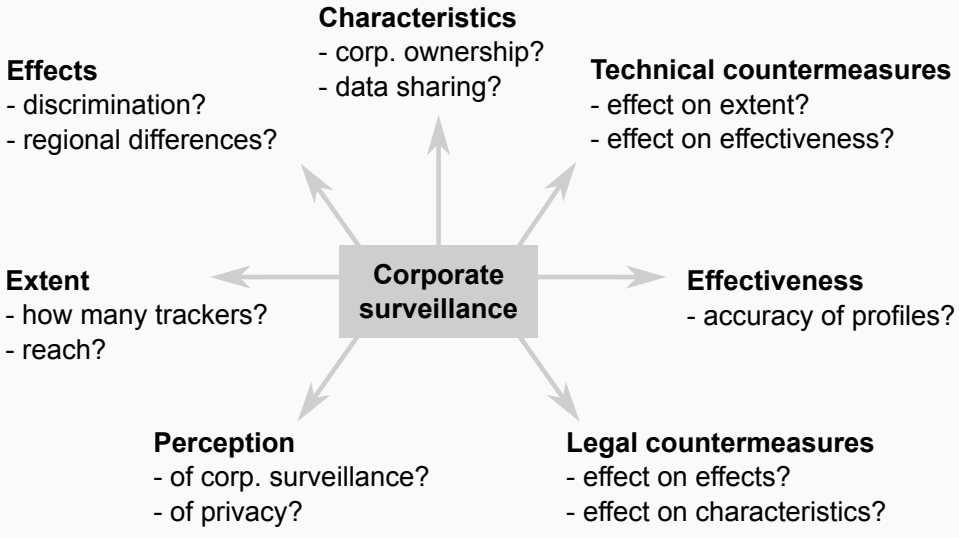
OVERVIEW

- Experiment: understand how changes to system's inputs change its outputs[1]
- Designing experiment means making detailed experimental plan:
  - Research questions
  - Templates for high-level study designs
  - Input variables
  - Output (response) variables
  - Challenges when studying black-box systems

[1] D. C. Montgomery, *Design and Analysis of Experiments*, 9 edition. Hoboken, NJ: Wiley, Apr. 2019.

Research questions

WHAT QUESTIONS CAN TRANSPARENCY RESEARCH ANSWER?



EXAMPLE RESEARCH QUESTIONS

- Characteristics and effects of corporate surveillance
  - Does price discrimination, facilitated by personal information, exist on the Internet?
  - How much do advertisers pay to reach a user?
- Extent of corporate surveillance
  - How many EU Facebook users have been assigned sensitive ad preferences?
  - How often do real users receive personalized search results?
- Effectiveness of corporate surveillance
  - What information is contained in advertising interest profiles, and are these profiles accurate?
- Effectiveness of countermeasures
  - To what extent do ad blockers and tracker blockers reduce exposure to tracking?
- Effectiveness of regulations and feedback mechanisms
  - Does the service adhere to its own privacy policy?
  - Did the GDPR have an effect on privacy policies or the presence of third parties on EU websites?

Study designs

STUDY DESIGNS

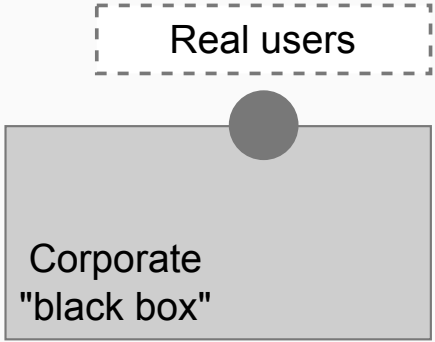
- Audit study: historically, a field experiment aimed at detecting discrimination[2]
- If designed and executed carefully: admissible as evidence in lawsuits
- Typical goal: detect whether information can flow from specific inputs to specific outputs

[2] C. Sandvig, K. Hamilton, K. Karahalios, *et al.*, "Auditing Algorithms: Research Methods for Detecting Discrimination on Internet Platforms," in *Data and Discrimination: Converting Critical Concerns into Productive Inquiry*, Seattle, WA, USA, May 2014, p. 23.

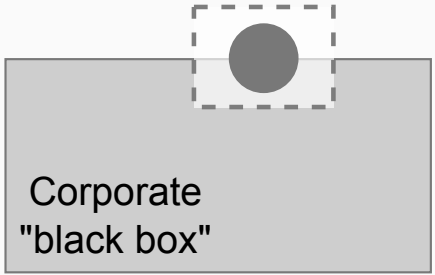
- Researcher inspects code used by the system
- Can allow definitive answers to research questions
- But code normally not available to researchers
  - Exceptions: JavaScript, mobile apps
- Code may not be enough: e.g., code+data for systems that use machine learning

Dashed line: subject of study
Dark circle: possible interactions with the corporate black-box system

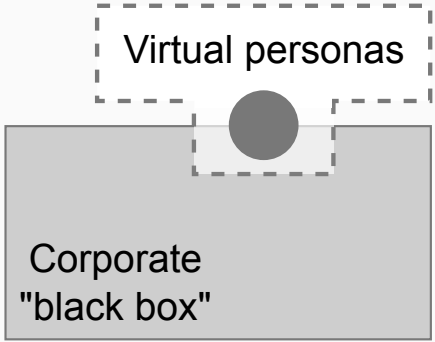
- Researcher asks users about their experiences with the system (surveys, interviews)
- Useful to study perceptions of corporate surveillance
- Challenges:
  - Unreliability of human memory
  - Cognitive biases



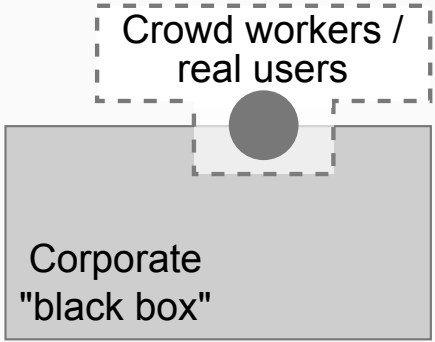
- Researcher interacts with system directly and repeatedly
- Observed responses allow inferences about system behavior
- Challenges:
  - Systems may detect automated requests and block or treat differently
  - May violate terms of service



- Researcher designs *virtual personas* to interact with the system
- Can recreate realistic user interactions
- Good control over input variables
- Allows randomized experiments
- Challenges:
  - May violate terms of service



- Researcher asks real userse to interact with the system
- Distribution of tasks facilitated by crowdsourcing platforms: Amazon Mechanical Turk, Prolific Academic, CrowdFlower
- Collected data based on real user profiles
- Challenges:
  - Large-scale experiments can have high monetary cost
  - More noise factors compared to sockpuppet studies



RANDOMIZED CONTROLLED TRIALS[3]

- Goal: check whether factors have an **effect** on behavior of a black-box system
- Each **experimental unit** is randomly assigned to receive either **control treatment** or **experimental treatment**
- In experimental treatment condition, **experimental factor** is varied while other factors are kept constant
- Experimenter measures **response** of each experimental unit to determine size of effect

Effect	Use of user data for marketing?
Experimental unit	Instances of user browser
Experimental factor	User behavior
Constant factors	IP address, time of day, etc.
Response	Sequences of ads

[3] M. C. Tschantz, A. Datta, A. Datta, *et al.*, "A Methodology for Information Flow Experiments," in *2015 IEEE 28th Computer Security Foundations Symposium*, Verona, Italy: IEEE, Jul. 2015, pp. 554–568. DOI: *10.1109/CSF.2015.40*.

LONGITUDINAL STUDIES

- Goal: understand long-term effects, evolution over time, changes before/after events
- In principle, same research questions and study designs, e.g.:
  - How did the prevalence and complexity of trackers evolve?[4]
  - How did the reach and penetration of trackers evolve?[5]
- Additional considerations:
  - Time period? (days/months/years)
  - Repeated data collection at specified times/intervals?
  - Sources with historical data?

[4] 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.

[5] 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.

Input variables / Experimental factors

CATEGORIES OF EXPERIMENTAL FACTORS

- User behavior
  - Browsing sequence to indicate user's interests
  - Search terms
  - Browser, browser extensions, operating system
  - Vantage points (geo-location, residential/university/data center connection)
- Websites
  - Top-N websites
  - E-commerce sites, publishers
- Advertising
  - Targeting attributes

EXPERIMENTAL FACTORS: WEBSITES

- To make quantitative and significant statement about tracking on the web, we need to analyze a representative sample of websites
- Most common: sample from the Alexa toplist
  - Ranking of most popular websites
  - But ranking changes frequently: almost half of the top 1 million sites changes from one day to the next
  - Research results not reproducible!
- New approach: tranco list[6]
  - Combines three different rankings: Alexa, Umbrella, Majestic
  - Makes resulting list more resistant to manipulation
  - Allows to reproduce research results because lists are archived and downloadable

[6] V. L. Pochat, T. van Goethem, and W. Joosen, "Rigging Research Results by Manipulating Top Websites Rankings.," in *26th Annual Network and Distributed System Security Symposium*, San Diego, CA, USA: Internet Society, Feb. 2019. DOI: *10.14722/ndss.2019.23386*.

SAMPLING STRATEGIES FOR WEBSITES

- Stratified sampling
  - Aim: include lower-ranked websites as well as higher-ranked websites, but keep sample size manageable
  - Examples: top 500 sites + 500 sampled uniformly at random from top 1 million, top 100,000 plus 2,000 from each subsequent 100,000 bracket
- Sampling from categories
  - Aim: improve diversity of sampled websites or focus on specific types of websites
  - Categories from Alexa or other domain classification services (McAfee)
  - Examples: top sites from each Alexa top-level category, top sites from selection of *interesting* categories
- Sampling from countries
  - Aim: study specific regions, such as EU, or differences between regions

SAMPLING STRATEGIES FOR WEBSITES: SUBSITES

- Subsites: internal pages that are not a domain's landing page
- Subsites have more cookies, trackers[7], and ads[8]
- Sampling strategies for subsites
  - Crawl landing pages and extract internal links
  - Sample websites from social media shares
  - Hispar toplist includes subsites

[7] T. Urban, M. Degeling, T. Holz, *et al.*, "Beyond the Front Page:Measuring Third Party Dynamics in the Field," in *Proceedings of The Web Conference 2020*, ser. WWW '20, Taipei, Taiwan: ACM, Apr. 2020, pp. 1275–1286. DOI: *10.1145/3366423.3380203*.

[8] W. Aqeel, B. Chandrasekaran, A. Feldmann, *et al.*, "On Landing and Internal Web Pages: The Strange Case of Jekyll and Hyde in Web Performance Measurement," in *Proceedings of the ACM Internet Measurement Conference*, ser. IMC '20, Pittsburgh, PA, USA: Association for Computing Machinery, Oct. 2020, pp. 680–695. DOI: *10.1145/3419394.3423626*.

EXPERIMENTAL FACTORS: MOBILE APPS

- Most popular source: Google Play store
  - 27 categories
  - Common filters: most popular (by ranking or by #downloads), free, use of specific permissions, inclusion of ad/analytics libraries
  - Number of apps studied: varies widely from ~100 to tens of thousands
- Other app stores
  - AppChina, AnZhi, Mi.com
- AndroZoo dataset[9]
  - Millions of apps from several app stores
  - Supports reproducible studies

[9] K. Allix, T. F. Bissyandé, J. Klein, *et al.*, "AndroZoo: Collecting Millions of Android Apps for the Research Community," in *2016 IEEE/ACM 13th Working Conference on Mining Software Repositories (MSR)*, May 2016, pp. 468–471.

EXPERIMENTAL FACTORS: VANTAGE POINTS

- Vantage points matter
  - Laws and regulations differ between geographical locations (e.g., GDPR, CCPA)
  - Some ad networks only serve ads to residential IP addresses[10]
  - Censorship varies between residential and data center IP addresses[11]
- Residential IP addresses via Luminati[12]
  - Proxy service, traffic exits at residential IP address in specified country
  - Can be expensive: smallest package has minimum monthly commitment of $500
- Geolocations
  - Via VPNs, university collaborations, cloud providers

[10] P. Vadrevu and R. Perdisci, "What You See is NOT What You Get: Discovering and Tracking Social Engineering Attack Campaigns," in *Proceedings of the Internet Measurement Conference*, ser. IMC '19, Amsterdam, Netherlands: ACM, Oct. 2019, pp. 308–321. DOI: 10.1145/3355369.3355600.

[11] R. Ramesh, R. S. Raman, M. Bernhard, *et al.*, "Decentralized Control: A Case Study of Russia," in *Proceedings 2020 Network and Distributed System Security Symposium*, San Diego, CA: Internet Society, 2020. DOI: 10.14722/ndss.2020.23098.

[12] A. McDonald, M. Bernhard, L. Valenta, *et al.*, "403 Forbidden: A Global View of CDN Geoblocking," in *Proceedings of the Internet Measurement Conference 2018*, ser. IMC '18, Boston, MA, USA: ACM, 2018, pp. 218–230. DOI: 10.1145/3278532.3278552.

VIRTUAL PERSONAS

- Aim: create a record in the browser and in the servers of web entities that mimic a real user, including demographics and interests
- Choosing interest profiles for personas
  - Select categories, e.g. from Google Adwords, Alexa
  - One persona per category
- Training personas: induce interests by visiting specific websites
  - List of related websites in Google Adwords Planner[13]
  - List of websites in Alexa category
  - Google search for relevant keywords[14]
- Control pages: websites to collect response variables
  - Weather, news

[13] J. M. Carrascosa, J. Mikians, R. Cuevas, *et al.*, "I Always Feel Like Somebody's Watching Me: Measuring Online Behavioural Advertising," in *Proceedings of the 11th ACM Conference on Emerging Networking Experiments and Technologies*, ser. CoNEXT '15, Heidelberg, Germany: ACM, 2015, 13:1–13:13. DOI: 10.1145/2716281.2836098.

[14] K. Solomos, P. Ilia, S. Ioannidis, *et al.*, "Talon: An Automated Framework for Cross-Device Tracking Detection," in *22nd International Symposium on Research in Attacks, Intrusions and Defenses (RAID 2019)*, Beijing, China: USENIX, Sep. 2019. arXiv: 1812.11393.

REPRESENTATIVE SAMPLES

- In studies with real users:
  - make sure that characteristics of participants are similar to characteristics of the underlying population
  - make sure that number of participants (sample size) is large enough to estimate their characteristics
- Underlying populations: all users of a social network, all users of a website, etc.
- Characteristics: often demographics (e.g., age, gender, education, income, location)
- Avoid convenience samples: students at the local university, social circle of researchers, users who self-select to install the researchers' app
- In all cases: report on sample characteristics

Response variables

RESPONSE VARIABLES (WEB)

- Highly dependent on research questions and study design
- Network communication
  - Properties of HTTP requests/responses, cookies, certificates, properties of HTML/JavaScript
- Advertising
  - Ad explanations, properties of ads, ad audience size estimate
- Search
  - Ranking of search results, autocomplete suggestions, product recommendations, prices, product ratings

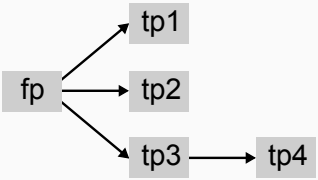
RESOURCE INCLUSION

- Two ways to record which resources are included in a website: HTTP requests and DOM tree
- But: no information where resources were included from
  - HTTP requests: referer header is first-party website, even if third-party JavaScript included a resource
  - DOM trees: can be modified dynamically, so parent/child relationship in DOM does not always correspond to resource inclusion

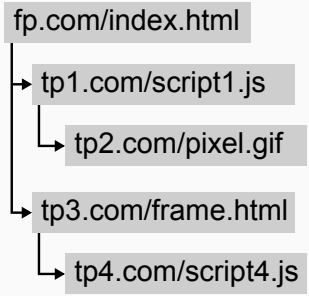
REQUEST TREES AND INCLUSION TREES

```
<script src="tp1.com/script1.js"></script>
<!-- pixel is dynamically inserted by script1.js -->
<img src="tp2.com/pixel.gif"/>
<iframe src="tp3.com/frame.html">
<script src="tp4.com/script4.js"></script>
</iframe>
```

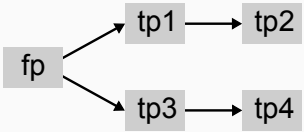
Request tree:



Inclusion tree:



Inclusion graph:



CONSTRUCTING INCLUSION TREES

- Recording of HTTP requests and DOM tree is not enough
- Instead, need to track static and dynamic resource inclusions
- Two options:
  - Modify browser (HTML parsing engine, extension engine)[15]
  - Use Chrome Debugging Protocol[16]
- Inclusion graph:
  - Union of all inclusion trees in a dataset
  - Shows relationship between all first parties and all third parties in the dataset
  - Edge weights can indicate number of resource inclusions

[15] M. A. Bashir, S. Arshad, C. Wilson, *et al.*, "Tracing Information Flows Between Ad Exchanges Using Retargeted Ads," in *25th USENIX Security Symposium*, Austin, TX, USA: USENIX, Aug. 2016, p. 17.

[16] M. A. Bashir and C. Wilson, "Diffusion of User Tracking Data in the Online Advertising Ecosystem," *Proceedings on Privacy Enhancing Technologies*, vol. 2018, no. 4, pp. 85–103, Oct. 2018. DOI: *10.1515/popets-2018-0033*.

RESPONSE VARIABLES (MOBILE)

- Network communication
- Static properties of apps: analysis of app binary without execution
  - Permissions
  - Use of third-party libraries
  - API calls
- Dynamic behavior of apps: monitoring app execution, possibly in controlled environment
  - Local files generated by app
  - Stack traces
  - Information flows (taint tracking)

RESPONSE VARIABLES (REAL USERS)

- Demographics: gender, age, ethnicity, income level, educational attainment, marital status, number of children, employment, and residence location
- User identifiers: international mobile subscriber identity (IMSI), IMEI, SIM number, Android serial number, Android advertising ID, phone number, MAC address, IP address, device serial number, Wi-Fi SSID
- User attributes and preferences

Challenges for experimental design

SCOPING

- Which black box should be studied?
- Which aspects of this black box should be studied?
- What is known about entities/components in the black box?
- Which interaction points are available?
- What data flows into / out of the black box?
- Which factors influence behavior of the black box?

OTHER CHALLENGES

- Assumption of independence and identical distribution do not hold
- Noise factors, e.g.
  - updates to the studied web service
  - distributed infrastructure
  - unknown components/behaviors inside the black box
- System under study may attempt to evade being studied, e.g. Facebook's ads and ad explanations[17]
- Profile contamination: the act of browsing websites for measurement purposes can change the profile information held by the server and thus alter measurement[18]

[17] J. B. Merrill and A. Tobin, "Facebook Moves to Block Ad Transparency Tools —…," *ProPublica*, Jan. 2019.

[18] P. Barford, I. Canadi, D. Krushevskaja, *et al.*, "Adscape: Harvesting and Analyzing Online Display Ads," in *Proceedings of the 23rd International Conference on World Wide Web*, ser. WWW '14, Seoul, Korea: ACM, 2014, pp. 597–608. DOI: 10.1145/2566486.2567992.

CONTROLLING NOISE FACTORS

- Many different strategies, depending on the specific experiment
- Examples:
  - Distributed infrastructure => use static DNS entries, send all requests from same subnet
  - Changes to system under study => run experiments in lock-step, synchronize measurement time
  - Use randomized controlled trials to replicate measurements
  - Include control accounts to measure baseline noise level
- Constant factors also influence results, e.g. vantage point:
  - Residential vs academic vs data center IP address
  - Geographical location of vantage point

Summary

EXPERIMENT DESIGN

- Research questions
- Study designs
- Randomized controlled trials
- Input variables
- Output (response) variables
- Scoping & controlling noise factors

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