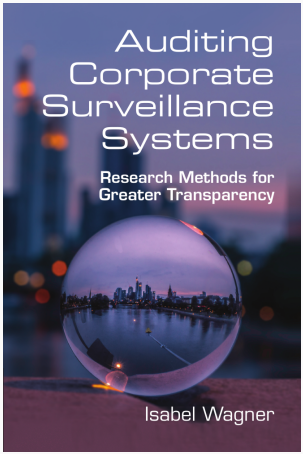
RESULTS FROM TRANSPARENCY RESEARCH

WEB SERVICES

Isabel Wagner

De Montfort University



Book design ©2022
by Cambridge University Press

OVERVIEW

- Web access
- Privacy policies and cookie notices
- Search
- Social networks
- Pricing
- Ratings and rankings
- Browser extensions
- Malicious web services

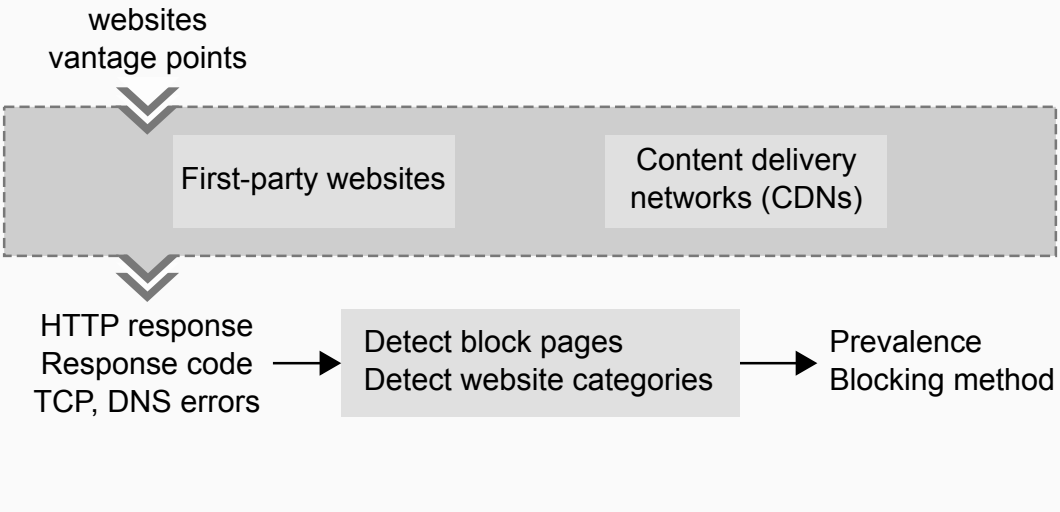
Web access

SERVER-SIDE BLOCKING

- Motivations for blocking users from accessing websites[1]
  - Block EU users to avoid having to comply with GDPR
  - Block regions due to economic sanctions
  - Block regions due to local third-party liability rules
  - Block due to hosting costs
  - Block due to security concerns
  - Block to prevent fraudulent traffic
- Extent of blocking?
- Technical mechanisms to perform blocking?

[1] M. C. Tschantz, S. Afroz, S. Sajid, *et al.*, "A Bestiary of Blocking: The Motivations and Modes behind Website Unavailability," in *8th USENIX Workshop on Free and Open Communications on the Internet (FOCI 18)*, Baltimore, MD, USA: USENIX, 2018, S. Afroz, M. C. Tschantz, S. Sajid, *et al.*, "Exploring Server-side Blocking of Regions," *arXiv:1805.11606 [cs]*, May 2018. arXiv: *1805.11606 [cs]*.

DESIGN FOR STUDYING SERVER-SIDE BLOCKING



EXTENT OF SERVER-SIDE BLOCKING

- Vantage points in Africa and Pakistan: <1% of websites geoblock
- Cloudflare-hosted websites: 0.6% perform country-based blocking, 10x more perform blocking for security reasons (IP ranges, browsers, CAPTCHAs)
- Vantage points in 177 countries:[2]
  - 4.4% of websites block at least one country
  - Shopping websites do most geoblocking
  - Iran, Syria, Sudan, Cuba are most frequently blocked countries
  - Median of 4 blocked websites per country
- Vantage point in Tor network: 16.7% of landing pages block Tor users, CAPTCHA for 40% of Google searches[3]

[2] 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*.

[3] I. Wagner, "A Machine Learning Approach to Detect Differential Treatment of Anonymous Users," in *European Symposium on Research in Computer Security*, Copenhagen, Denmark, Sep. 2022.

WIFI HOTSPOTS[4]

- Study behavior of 67 open hotspots in Montreal: traffic from captive portal and landing pages
- 40% collect PII via registration or social login (mandatory on 28%)
- 95% have trackers on captive portal, 7.4 trackers on average
- 38% set cookies (some with 20-year expiration) before user gives consent
- 15% fingerprint before user consent
- Data shared with third parties: MAC address (59%), PII (8%)

[4] S. Ali, T. Osman, M. Mannan, *et al.*, "On Privacy Risks of Public WiFi Captive Portals," in *Data Privacy Management, Cryptocurrencies and Blockchain Technology*, C. Pérez-Solà, G. Navarro-Arribas, A. Biryukov, *et al.*, Eds., ser. Lecture Notes in Computer Science, Cham: Springer International Publishing, 2019, pp. 80–98. DOI: *10.1007/978-3-030-31500-9_6*.

Privacy policies and cookie notices

COOKIE NOTICES POST-GDPR[7]

- Prevalence of cookie notices: 46% pre-GDPR, 63% post-GDPR
- Implementation of cookie notices
  - 2019: 15% of websites use third-party consent library, one-third supports centralized consent management (IAB Transparency and Consent Framework (TCF))
  - 2020: use of centralized consent management doubles every year[5], TCF banners on 6.2% of websites
  - TCF banners not necessarily GDPR compliant!
  - 6.2% store consent prior to user choice, 6.8% do not provide way to opt out, 46.5% pre-select consent, 5.3% store positive consent regardless of user choice[6]
  - Denying consent takes longer: median of 6.7s vs. 3.2s for giving consent

[5] M. Hils, D. W. Woods, and R. Böhme, "Measuring the Emergence of Consent Management on the Web," in *Proceedings of the ACM Internet Measurement Conference*, ser. IMC '20, Pittsburgh, PA, USA: Association for Computing Machinery, Oct. 2020, pp. 317–332. DOI: 10.1145/3419394.3423647.

[6] C. Matte, N. Bielova, and C. Santos, "Do Cookie Banners Respect my Choice?" In *2020 IEEE Symposium on Security and Privacy (SP)*, San Francisco, CA, USA: IEEE, May 2020, pp. 1612–1630. DOI: 10.1109/SP40000.2020.00076.

[7] M. Degeling, C. Utz, C. Lentzsch, *et al.*, "We Value Your Privacy ... Now Take Some Cookies: Measuring the GDPR's Impact on Web Privacy," in *Network and Distributed Systems Security (NDSS) Symposium*, San Diego, CA, USA: Internet Society, Feb. 2019. DOI: 10.14722/ndss.2019.23378.

PRIVACY POLICIES POST-GDPR

- 15% of websites newly introduced privacy policies
- 72% of websites updated their privacy policy
- Privacy policies became 25–40% longer[8]
- Opt-out choices for users: embedded as links only in 3% of policies[9]

[8] M. Degeling, C. Utz, C. Lentzsch, *et al.*, "We Value Your Privacy ... Now Take Some Cookies: Measuring the GDPR's Impact on Web Privacy," in *Network and Distributed Systems Security (NDSS) Symposium*, San Diego, CA, USA: Internet Society, Feb. 2019. DOI: 10.14722/ndss.2019.23378, T. Linden, R. Khandelwal, H. Harkous, *et al.*, "The Privacy Policy Landscape After the GDPR," *Proceedings on Privacy Enhancing Technologies*, vol. 2020, no. 1, pp. 47–64, Jan. 2020. DOI: 10.2478/popets-2020-0004.

[9] V. B. Kumar, R. Iyengar, N. Nisal, *et al.*, "Finding a Choice in a Haystack: Automatic Extraction of Opt-Out Statements from Privacy Policy Text," in *Proceedings of The Web Conference 2020*, ser. WWW '20, Taipei, Taiwan: Association for Computing Machinery, Apr. 2020, pp. 1943–1954. DOI: 10.1145/3366423.3380262.

CONSISTENCY OF PRIVACY POLICIES

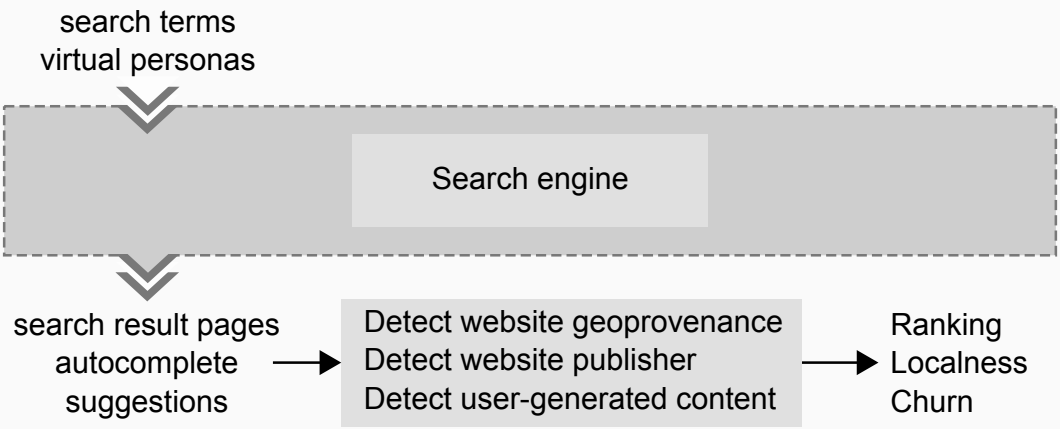
- Consistency with observed data flows[10]
  - The policies of 42% of apps disclose relevant data flows incorrectly or not at all
- Contradictory statements[11]
  - 14% have logical contradictions, 17% have contradictions or narrowing definitions
  - 60% of policies contain negations (harder to understand)
- Consistency with requested permissions, API use[12]
  - 88% of apps perform data practice that would have to be disclosed in a policy
  - Only 50% of apps have a privacy policy
  - Median 3 compliance issues per app
  - 12% of apps have compliance issues related to location

[10] B. Andow, S. Y. Mahmud, J. Whitaker, *et al.*, "Actions Speak Louder than Words: Entity-Sensitive Privacy Policy and Data Flow Analysis with PoliCheck," in *29th USENIX Security Symposium (USENIX Security 20)*, online: USENIX, Aug. 2020, pp. 985–1002.

[11] B. Andow, S. Y. Mahmud, W. Wang, *et al.*, "PolicyLint: Investigating Internal Privacy Policy Contradictions on Google Play," in *28th USENIX Security Symposium (USENIX Security 19)*, Santa Clara, CA, USA, Aug. 2019, pp. 585–602.

[12] S. Zimmeck, P. Story, D. Smullen, *et al.*, "MAPS: Scaling Privacy Compliance Analysis to a Million Apps," *Proceedings on Privacy Enhancing Technologies*, vol. 2019, no. 3, pp. 66–86, Jul. 2019. DOI: 10.2478/popets-2019-0037.

Search


search terms
virtual personas
Search engine
search result pages
autocomplete
suggestions
Detect website geoprovenance
Detect website publisher
Detect user-generated content
Ranking
Localness
Churn

PARTISANSHIP OF TEXT SNIPPETS IN SEARCH RESULTS[13]

- Text snippets influence which search results users click on
- For political queries, are they partisan?
  - Partisan: associated with a political leaning (left/right)
  - E.g., through use of words: “gun rights” (right) vs “gun violence” (left)
- 50% of snippets amplify partisanship (snippet is more partisan than its website)
- 20% have inverse partisanship
- Likely cause: not Google’s snippet algorithm, but practice of starting news articles with strong language

[13] D. Hu, S. Jiang, R. E. Robertson, *et al.*, “Auditing the Partisanship of Google Search Snippets,” in *The World Wide Web Conference*, ser. WWW ’19, San Francisco, CA, USA: ACM, 2019, pp. 693–704. DOI: *10.1145/3308558.3313654*.

GOOGLE'S TOP STORIES ALGORITHM

- Top stories: 3 news articles are highlighted when searching for trending stories
- Which publishers are selected to provide these stories?[14]
  - Less than 20 publishers provide more than half of stories
  - Top 20% of publishers provide 90% of stories
- Do publishers differ between first/second/third spot?
  - One-third of publishers appears at least once in first spot
  - Two-thirds of publishers appear at least once in third spot
  - Google may choose lesser-known publishers for third spot

[14] E. Lurie and E. Mustafaraj, "Opening Up the Black Box: Auditing Google's Top Stories Algorithm," in *Proceedings of the 32nd International Florida Artificial Intelligence Research Society Conference*, vol. 32, Sarasota, Florida, USA: AAAI Press, May 2019, pp. 376–382, D. Trielli and N. Diakopoulos, "Search As News Curator: The Role of Google in Shaping Attention to News Information," in *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, ser. CHI '19, Glasgow, Scotland Uk: ACM, 2019, 453:1–453:15. DOI: *10.1145/3290605.3300683*.

AUTOCOMPLETE SUGGESTIONS[15]

- Important because they influence which search terms users use
- Differences between Google and Bing:
  - More negative emotions in Bing suggestions
  - More social media references in Google suggestions (10% vs 4.6%)
  - Social media references on Google mostly point to YouTube, i.e., Google's autocomplete algorithm favors Google's own products
- Suggestions change frequently: most suggestions present for <10 days in 10-week dataset
- Suggestions higher in the list are more stable, perhaps to combat inaccurate or misleading suggestions

[15] R. E. Robertson, S. Jiang, D. Lazer, *et al.*, "Auditing Autocomplete: Suggestion Networks and Recursive Algorithm Interrogation," in *Proceedings of the 10th ACM Conference on Web Science - WebSci '19*, Boston, Massachusetts, USA: ACM Press, 2019, pp. 235–244. DOI: *10.1145/3292522.3326047*.

How does Google personalize search results?

- Personalization of search results:
  - What data does Google use for personalization?
  - How big is the effect of personalization?
- 2013: 11% of results are personalized[16]
  - More towards bottom of list
  - Mostly political, news, local results
  - Personalization based on logged-in account and IP
- 2018: Search result pages from autocomplete suggestions are personalized[17]
  - For political searches: more personalization if user has weak political preference
  - Logged-in users receive 19% more personalization

[16] A. Hannak, P. Sapiezynski, A. Molavi Kakhki, *et al.*, "Measuring Personalization of Web Search," in *Proceedings of the 22nd International Conference on World Wide Web*, ser. WWW '13, Rio de Janeiro, Brazil: ACM, 2013, pp. 527–538. DOI: *10.1145/2488388.2488435*.

[17] R. E. Robertson, D. Lazer, and C. Wilson, "Auditing the Personalization and Composition of Politically-Related Search Engine Results Pages," in *Proceedings of the 2018 World Wide Web Conference*, ser. WWW '18, Lyon, France: International World Wide Web Conferences Steering Committee, 2018, pp. 955–965. DOI: *10.1145/3178876.3186143*.

NEWS CURATION

- Google News[18]
  - Personalization reinforces partisanship of virtual personas
  - Search result list differs by 3.2% depending on persona, edit distance of 8.4 in 100 results
  - Larger differences toward bottom of list
- Apple News[19]
  - Top stories are human-curated, trending stories algorithmically curated
  - No personalization found
  - Concentration of publishers: top 3 publishers provide 23% of stories for human-curated, compared to 45% for algorithmically-curated
  - Selection of topics: human-curated focuses on substantive policy issues, international news; algorithmically-curated features celebrity and sensational news

[18] H. Le, R. Maragh, B. Ekdale, *et al.*, "Measuring Political Personalization of Google News Search," in *The World Wide Web Conference*, ser. WWW '19, San Francisco, CA, USA: ACM, 2019, pp. 2957–2963. DOI: *10.1145/3308558.3313682*.

[19] J. Bandy and N. Diakopoulos, "Auditing News Curation Systems: A Case Study Examining Algorithmic and Editorial Logic in Apple News," *arXiv:1908.00456 [cs]*, Aug. 2019. arXiv: *1908.00456 [cs]*.

How 'local' are search results on localized Google?[20]

- Localized search engines:
  - When searching for local information, how many of the search results are local sources?
  - Search for city names, observe where search results are located
  - Localness indicator: ratio of locally-produced sources in search results to all search results
- Wealthy countries have more local results
- Places in Global South tend to be represented by sources from Global North
  - US-generated content represents half of first-page search results for 61 countries
  - Digital hegemony: content producers in few countries determine what information is available about other countries

[20] A. Ballatore, M. Graham, and S. Sen, "Digital Hegemonies: The Localness of Search Engine Results," *Annals of the American Association of Geographers*, vol. 107, no. 5, pp. 1194–1215, Sep. 2017. DOI: *10.1080/24694452.2017.1308240*.

POLITICAL BIASES IN SEARCH RESULTS

- Twitter[21]
  - Bias of tweets: slightly left-leaning
  - Bias of Twitter's ranked list of tweets (when searching for political candidates): more left-leaning for candidates from left, more right-leaning for candidates from right
  - Ranking algorithm amplifies bias
- Google[22]
  - No difference in weighted bias (top results weighted higher), indicates absence of filter bubble
  - Results in Google's twitter-card: more right-leaning than other components of search result page (SERP)
  - Lower-ranked search results more left-leaning
  - Ranking algorithm shifts SERP to the right, magnitude depends on search term

[21] J. Kulshrestha, M. Eslami, J. Messias, *et al.*, "Quantifying Search Bias: Investigating Sources of Bias for Political Searches in Social Media," in *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing*, ser. CSCW '17, Portland, Oregon, USA: ACM, 2017, pp. 417–432. DOI: 10.1145/2998181.2998321.

[22] R. E. Robertson, S. Jiang, K. Joseph, *et al.*, "Auditing Partisan Audience Bias Within Google Search," *Proc. ACM Hum.-Comput. Interact.*, vol. 2, no. CSCW,

SEARCH ENGINE MANIPULATION EFFECT[23]

- Can individuals be influenced by biased search rankings?
- Lab study: voting preferences of undecided voters can shift by 20% or more
  - Foreign election with two candidates
  - Three groups of participants, receive search results to make informed choice
  - Three sets of search results: unbiased + 2x biased towards a candidate
- Most participants were not aware that search ranking was biased
- If 10% of voters are undecided: effect could shift election outcome by 2.5%

[23] R. Epstein and R. E. Robertson, "The search engine manipulation effect (SEME) and its possible impact on the outcomes of elections," *Proceedings of the National Academy of Sciences*, vol. 112, no. 33, E4512–E4521, Aug. 2015. DOI: *10.1073/pnas.1419828112*.

Social networks

EMOTIONAL CONTAGION[24]

- Facebook experiment with 689,000 participants
  - No informed consent
  - Facebook claims their data use policy covers experimenting on their users
  - First classify emotions in user posts (positive/negative)
  - Then show some users more positive posts, others more negative posts
  - Observe emotions in user's subsequent posts
  - Users who received fewer positive posts: use 0.1% fewer positive words, 0.04% more negative words
  - Users who received fewer negative posts: use 0.07% fewer negative words, 0.06% more positive words
- Facebook has the power to influence how its users feel, without users noticing
- Are they using this power? Demonstrates why it is essential that we audit social networks!

[24] A. D. I. Kramer, J. E. Guillory, and J. T. Hancock, "Experimental evidence of massive-scale emotional contagion through social networks," *Proceedings of the National Academy of Sciences*, vol. 111, no. 24, pp. 8788–8790, Jun. 2014. DOI: *10.1073/pnas.1320040111*.

POLITICAL MOBILIZATION[25]

- During 2010 US congressional and 2012 US presidential elections: Facebook displayed messages to encourage people to vote
- Three groups of users:
  - Informational message: voting day, nearest polling place
  - Social message: information + pictures of friends who have already voted
  - No message
- Social message group 0.39% and 0.24% more likely to vote
- 200,000+ additional votes, may influence election outcome
- Facebook has the power to use political mobilization in discriminatory way: show social message only to users who are expected to vote a certain way
- Are they using this power?

[25] R. M. Bond, C. J. Fariss, J. J. Jones, *et al.*, "A 61-million-person experiment in social influence and political mobilization," *Nature*, vol. 489, no. 7415, pp. 295–298, Sep. 2012. DOI: *10.1038/nature11421*, J. J. Jones, R. M. Bond, E. Bakshy, *et al.*, "Social influence and political mobilization: Further evidence from a randomized experiment in the 2012 U.S. presidential election," *PLOS ONE*, vol. 12, no. 4, e0173851, Apr. 2017. DOI: *10.1371/journal.pone.0173851*.

BIASES IN FACEBOOK NEWS FEED

- Facebook: “the order in which users see stories in the News Feed depends on many factors, including how often the viewer visits Facebook, how much they interact with certain friends, and how often users have clicked on links to certain websites in News Feed in the past”[26]
- 2018 Italian election:[27]
  - Virtual personas with different preferences: each persona follows all political parties, preferences induced by liking and commenting
  - Observe order of posts in news feed
  - Top position in news feed: strongly biased towards persona’s political leaning
  - News feed for “undecided” persona: similar to feed for right-leaning persona

[26] E. Bakshy, S. Messing, and L. A. Adamic, “Exposure to ideologically diverse news and opinion on Facebook,” *Science*, vol. 348, no. 6239, pp. 1130–1132, Jun. 2015. DOI: *10.1126/science.aaa1160*.

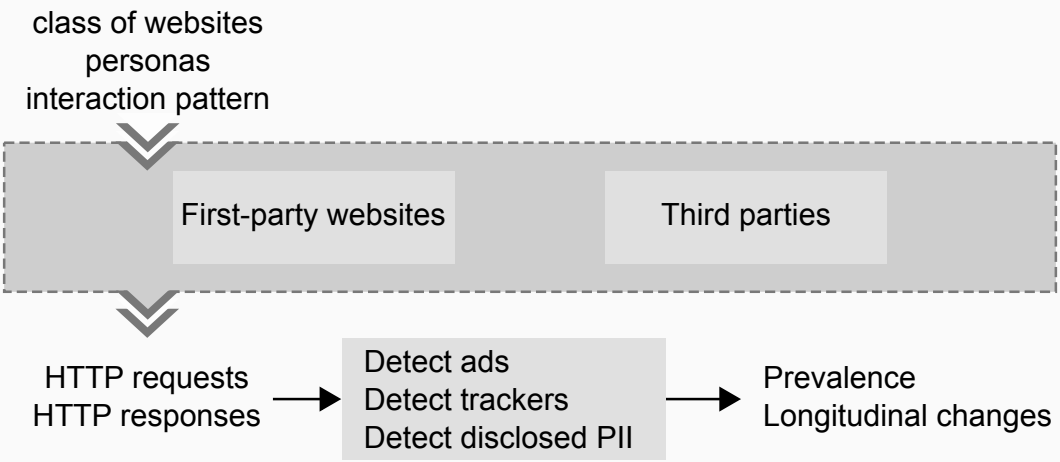
[27] E. Hargreaves, C. Agosti, D. Menasché, *et al.*, “Biases in the Facebook News Feed: A Case Study on the Italian Elections,” in *2018 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, Aug. 2018, pp. 806–812. DOI: *10.1109/ASONAM.2018.8508659*, E. Hargreaves, C. Agosti, D. Menasché, *et al.*, “Fairness in online social network timelines: Measurements, models and mechanism design,” *Performance Evaluation*, vol. 129, pp. 15–39, Feb. 2019. DOI: *10.1016/j.peva.2018.09.009*.

MALICIOUS LINKS ON SOCIAL PLATFORMS[28]

- Link sharing: social platform generates preview that is shown to users instead of URL
- Manipulation of link previews:
  - Attackers can exploit how link previews are crafted so that link preview hides the malicious target
  - Link previews are updated infrequently (every two weeks)
  - Attacker has a long time to run malicious campaigns
- 90% of social media and instant messaging sites do not use any countermeasures against sharing of malicious links

[28] G. Stivala and G. Pellegrino, "Deceptive Previews: A Study of the Link Preview Trustworthiness in Social Platforms," in *Proceedings 2020 Network and Distributed System Security Symposium*, San Diego, CA: Internet Society, 2020. DOI: 10.14722/ndss.2020.24252.

Pricing and ranking


class of websites
personas
interaction pattern
First-party websites
Third parties
HTTP requests
HTTP responses
Detect ads
Detect trackers
Detect disclosed PII
Prevalence
Longitudinal changes

PRICE DISCRIMINATION

- Cross-border price discrimination[29]
  - User-triggered price measurement from multiple vantage points across globe
  - 3.8% of retailers vary prices based on country
  - Variations up to factor of 7
  - Within-country variations attributable to A/B testing
- Price discrimination and steering[30]
  - Evidence for member-only prices, A/B testing
  - Personalization for mobile devices: expensive products first for some, price reduction for others
  - Some personalization based on purchase history

[29] C. Iordanou, C. Soriente, M. Sirivianos, *et al.*, "Who is Fiddling with Prices?: Building and Deploying a Watchdog Service for E-commerce," in *Proceedings of the Conference of the ACM Special Interest Group on Data Communication*, ser. SIGCOMM '17, Los Angeles, CA, USA: ACM, 2017, pp. 376–389. DOI: *10.1145/3098822.3098850*.

[30] A. Hannak, G. Soeller, D. Lazer, *et al.*, "Measuring Price Discrimination and Steering on E-commerce Web Sites," in *Proceedings of the 2014 Conference on Internet Measurement Conference*, ser. IMC '14, Vancouver, BC, Canada: ACM, 2014, pp. 305–318. DOI: *10.1145/2663716.2663744*.

ALGORITHMIC PRICING[31]

- Prices are set by an algorithm, not a human merchant
- Detecting algorithmic pricing on Amazon Marketplace
  - Assume that algorithmic prices are set in response to other sellers' prices
  - Observe time series of prices for specific products: seller's price, lowest price, Amazon's price, second-lowest price
  - If time series are similar, prices change at similar times
  - Pricing is algorithmic if Spearman's Rank Correlation $\rho \geq 0.7$ and $p \leq 0.05$
- Algorithmic pricing on Amazon Marketplace
  - 2.4% of sellers use it
  - 70% of algorithmic sellers within $1 of lowest price, 40% of Amazon's price
  - Algorithmic sellers active for longer, but sell fewer products than other sellers

[31] L. Chen, A. Mislove, and C. Wilson, "An Empirical Analysis of Algorithmic Pricing on Amazon Marketplace," in *Proceedings of the 25th International Conference on World Wide Web*, ser. WWW '16, Montréal, Québec, Canada: International World Wide Web Conferences Steering Committee, 2016, pp. 1339–1349. DOI: 10.1145/2872427.2883089.

DARK PATTERNS

- Patterns in user interface design, intended to steer users towards decisions that are not favorable to the user, but to the site employing the dark pattern[32]
- On e-commerce sites:[33]
  - 11% of shopping sites use dark patterns
  - Specialized third parties provide turnkey solutions for dark patterns
  - Most common patterns:
  - Low-stock messages (impression of scarcity)
  - Countdown timers (deal/discount will expire, induce time pressure)
  - Social activity messages (show other purchasers)

[32] C. Bösch, B. Erb, F. Kargl, *et al.*, "Tales from the Dark Side: Privacy Dark Strategies and Privacy Dark Patterns," *Proceedings on Privacy Enhancing Technologies*, vol. 2016, no. 4, pp. 237–254, Oct. 2016. DOI: *10.1515/popets-2016-0038*.

[33] A. Mathur, G. Acar, M. J. Friedman, *et al.*, "Dark Patterns at Scale: Findings from a Crawl of 11K Shopping Websites," *Proceedings of the ACM on Human-Computer Interaction*, vol. 3, no. CSCW, 81:1–81:32, Nov. 2019. DOI: *10.1145/3359183*.

BIASES IN HOTEL RANKING ALGORITHMS[34]

- On hotel booking sites, does the rank of a hotel depend on its price on other booking sites?
  - Price parity clauses: often used by booking sites to require hotels to make price on booking site match the lowest price elsewhere
  - Prohibited by EU competition regulators (2013–2018)
  - Ranking penalty a way to punish hotels with lower prices elsewhere?
- Sample of 18,000 hotels, 25% had lower prices on hotel's own website, 15–20% on competitor booking site
- 10% lower price elsewhere reduced hotel ranking as if user rating was lowered by 0.3 points (out of 10)
- Ranking penalty is real
- Search quality for consumers is reduced

[34] M. Hunold, R. Kesler, and U. Laitenberger, "Hotel rankings of online travel agents, channel pricing and consumer protection," DICE Discussion Paper, Working Paper 300, 2018.

BIASES IN RÉSUMÉ SEARCH WEBSITES[35]

- Recruiters search for job candidates
- Does résumé ranking algorithm have gender bias?
- No evidence for direct discrimination, i.e., gender is not input to ranking algorithm
- Individual fairness: candidates with similar features should be at similar ranks
  - At rank 30: females are 1.4 ranks below males
  - Statistically significant, but small effect
- Group fairness: distribution of ranks should be similar among men and women
  - 10% of combinations (job title/city) showed group unfairness
  - Sometimes men were the disadvantaged group, depending on job title

[35] L. Chen, R. Ma, A. Hannák, *et al.*, "Investigating the Impact of Gender on Rank in Resume Search Engines," in *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, Montréal, Québec, Canada: ACM, Apr. 2018, p. 651. DOI: 10.1145/3173574.3174225.

Browser extensions

USER DATA LEAKS FROM BROWSER EXTENSIONS

- Extensions may leak data to third parties
  - 2017: 6.3% of top 10,000 Chrome extensions leaked data
  - Many accidental through Referer header, but also: browsing history, search queries, form entries[36]
  - 2017: 1.9% of extensions tracked users + browsing history[37]
  - 2018: 2.1% of top 178,000 Chrome extension leaked sensitive information
- Some extensions attempt to evade detection: encryption, WebSockets
- Top 10 leaking extensions: 60 million users[38]

[36] O. Starov and N. Nikiforakis, "Extended Tracking Powers: Measuring the Privacy Diffusion Enabled by Browser Extensions," in *Proceedings of the 26th International Conference on World Wide Web*, ser. WWW '17, Perth, Australia: International World Wide Web Conferences Steering Committee, 2017, pp. 1481–1490. DOI: *10.1145/3038912.3052596*.

[37] M. Weissbacher, E. Mariconti, G. Suarez-Tangil, *et al.*, "Ex-Ray: Detection of History-Leaking Browser Extensions," in *Proceedings of the 33rd Annual Computer Security Applications Conference*, ser. ACSAC 2017, Orlando, FL, USA: ACM, 2017, pp. 590–602. DOI: *10.1145/3134600.3134632*.

[38] Q. Chen and A. Kapravelos, "Mystique: Uncovering Information Leakage from Browser Extensions," in *Proceedings of the 2018 ACM SIGSAC Conference on Computer and Communications Security*, ser. CCS '18, Toronto, Canada: ACM, 2018, pp. 1687–1700. DOI: *10.1145/3243734.3243823*.

Other web services

CLICK INTERCEPTION OR CLICKJACKING[39]

- Make users perform unintended clicks to generate ad revenue or distribute malicious content
- Techniques: e.g., invisible iFrame over benign-looking button
- Detecting clickjacking:
  - Browser extension monitors DOM, particularly HTML anchors, creation/modification of hyperlinks, JS event listeners
- On Alexa top 250,000: 437 unique third-party clickjacking scripts
- Clickjacking performed on 613 websites
- Techniques needed to ensure link integrity, limit privileges of third-party JS

[39] M. Zhang, W. Meng, S. Lee, *et al.*, "All Your Clicks Belong to Me: Investigating Click Interception on the Web," in *28th USENIX Security Symposium (USENIX Security 19)*, Santa Clara, CA, USA: USENIX, Aug. 2019, pp. 941–957.

CLOAKING[40]

- Cloaking: websites appear benign to search engines, but serve malicious content to humans
- Detection of cloaking:
  - Compare link titles and text snippets of search results with real website contents
- Cloaking more common for high-risk search terms:
  - Luxury products, weight loss products
  - 11.7% of top 100 search results and 4.9% of ads cloaked against Google crawler
- Client-side techniques needed to detect cloaking

[40] L. Invernizzi, K. Thomas, A. Kapravelos, *et al.*, "Cloak of Visibility: Detecting When Machines Browse a Different Web," in *2016 IEEE Symposium on Security and Privacy (SP)*, San Jose, CA, USA: IEEE, May 2016, pp. 743–758. DOI: 10.1109/SP.2016.50.

PAYWALLS[41]

- Alternative to ad-based business model: users who have paid subscription can log in and consume content freely
  - Soft paywall: free access to some articles (e.g., each month)
  - Hard paywall: subscription needed for all content
- Detecting paywalls:
  - Heuristics: paywall recognition rules from browser extensions and Fanboy's enhanced tracking list
  - Machine learning: random forest classifier based on text+structural+visual features achieves 77% precision
- Prevalence of paywalls: doubles every ~6 months
  - 80% of paywalled sites are news sites
  - US and Australia have highest paywall rate
- But: even paywalled sites use trackers and advertising

[41] P. Papadopoulos, P. Snyder, D. Athanasakis, *et al.*, "Keeping out the Masses: Understanding the Popularity and Implications of Internet Paywalls," in *Proceedings of The Web Conference 2020*, ser. WWW '20, Taipei, Taiwan: ACM, Apr. 2020, pp. 1433–1444. DOI: *10.1145/3366423.3380217*.

MAP PERSONALIZATION[42]

- No world-wide consensus about country borders: border disputes sometimes persist for many years
- Google Maps and Bing Maps use personalization to display different borders, depending on user's country
- Three versions for disputed map tiles: two with different border versions, one that indicates the border dispute
- 2016: for 7 regions with disputed borders
  - Google Maps shows personalized border for 5
  - Bing Maps shows personalized border for 4
- Updates to personalized border tiles can be very quick: Google updated its map 1-2 months after Russian annexation of Crimea
  - Harder to archive historical map material
  - Promotes diverging world view

[42] G. Soeller, K. Karahalios, C. Sandvig, *et al.*, "MapWatch: Detecting and Monitoring International Border Personalization on Online Maps," in *Proceedings of the 25th International Conference on World Wide Web*, ser. WWW '16, Montréal, Québec, Canada: International World Wide Web Conferences Steering Committee, 2016, pp. 867–878. DOI: 10.1145/2872427.2883016.

Summary

RESULTS FOR WEB SERVICES

- Study designs:
  - Server-side blocking
  - Search engines
  - Generic web services
- Overview of results:
  - Access to the web
  - Cookie notices and privacy policies
  - Search engines
  - Social networks
  - Pricing and ranking algorithms
  - Browser extensions
  - Other web services

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