[52] Jeff Huang, Patrick O'Neil Meredith, and Grigore Rosu. 2014. Maximal sound predictive race detection with control flow abstraction. *ACM SIGPLAN Notices*, 49, 6, 337–348. DOI: 10.1145/2666356.2594315.

[53] Ilya Sergey. 2019. What does it mean for a program analysis to be sound? SIGPLAN Blog. Retrieved 10/03/2020 from https://blog.sigplan.org/2019/08/07/what-does-it-mean-for-a-program-analysis-to-be-sound/.

[54] Google Inc. 2021. Robots.txt introduction & guide | google search central. Google Developers. Retrieved 07/09/2021 from https://developers.google.com/search/docs/advanced/robots/intro.

[55] Network Advertising Initiative. 2020. FAQ | NAI: network advertising initiative. Retrieved 03/14/2021 from https://www.networkadvertising.org/faq.

[56] Network Advertising Initiative. 2021. NAI consumer opt out. Retrieved 04/15/2021 from https://optout.networkadvertising.org/.

[57] Costas Iordanou, Georgios Smaragdakis, Ingmar Poese, and Nikolaos Laoutaris. 2018. Tracing cross border web tracking. In *Proceedings of the Internet Measurement Conference 2018*. ACM, New York, NY, USA, 329–342. DOI: 10.1145/3278532.3278561.

[58] Haojian Jin, Minyi Liu, Kevan Dodhia, Yuanchun Li, Gaurav Srivastava, Matthew Fredrikson, Yuvraj Agarwal, and Jason I. Hong. 2018. Why Are They Collecting My Data?: Inferring the Purposes of Network Traffic in Mobile Apps. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 2, 4, 173. DOI: 10.1145/3287051.

[59] Haojian Jin, Tetsuya Sakai, and Koji Yatani. 2014. ReviewCollage: a mobile interface for direct comparison using online reviews. In *Proceedings of the 16th international conference on Human-computer interaction with mobile devices & services*. ACM, New York, NY, USA, 349–358. DOI: 10.1145/2628363.2628373.

[60] Garrett A. Johnson, Scott K. Shriver, and Shaoyin Du. 2020. Consumer Privacy Choice in Online Advertising: Who Opts Out and at What Cost to Industry? *Marketing Science*, 39, 1, 33–51. DOI: 10.1287/mksc.2019.1198.

[61] Jordan Jueckstock, Shaown Sarker, Peter Snyder, Aidan Beggs, Panagiotis Papadopoulos, Matteo Varvello, Ben Livshits, and Alexandros Kapravelos. 2021. Towards Realistic and Reproducible Web Crawl Measurements. In *Proceedings of the The Web Conference*.

[62] Daniel Jurafsky and James H. Martin. 2019. *Speech and Language Processing, 3rd edition*. (Third Edition draft edition).

[63] Arjaldo Karaj, Sam Macbeth, Rémi Berson, and Josep M. Pujol. 2019. WhoTracks.me: shedding light on the opaque world of online tracking. *arXiv:1804.08959 [cs]*. arXiv: 1804.08959.

[64] Saranga Komanduri, Richard Shay, Greg Norcie, Blase Ur, and Lorrie Faith Cranor. 2011. AdChoices? Compliance with Online Behavioral Advertising Notice and Choice Requirements. *I/S: A Journal of Law and Policy for the Information Society*, 7, 603.

[65] John Kurkowski. 2020. Tldextract. Retrieved 03/15/2021 from https://github.com/john-kurkowski/tldextract.

[66] Daniil Larionov, Artem Shelmanov, Elena Chistova, and Ivan Smirnov. 2019. Semantic Role Labeling with Pretrained Language Models for Known and Unknown Predicates. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing*. INCOMA Ltd., Varna, Bulgaria, 619–628. DOI: 10.26615/978-954-452-056-4_073.

[67] Richard Lawler. 2022. Google delays blocking third-party cookies again, now targeting late 2024. Retrieved 08/08/2022 from https://www.theverge.com/2022/7/27/23280905/google-chrome-cookies-privacy-sandbox-advertising.

[68] Victor Le Pochat, Tom Van Goethem, Samaneh Tajalizadehkhoob, Maciej Korczyński, and Wouter Joosen. 2019. Tranco: a research-oriented top sites ranking hardened against manipulation. In *Proceedings of the 26th Annual Network and Distributed System Security Symposium*. DOI: 10.14722/ndss.2019.23386.

[69] Adam Lerner, Anna Kornfeld Simpson, Tadayoshi Kohno, and Franziska Roesner. 2016. Internet jones and the raiders of the lost trackers: an archaeological study of web tracking from 1996 to 2016. In *25th USENIX Security Symposium*.

[70] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A Robustly Optimized BERT Pretraining Approach. *arXiv:1907.11692 [cs]*.

[71] Will Bontrager Software LLC. 2021. Linking without an 'a' tag. Retrieved 03/01/2021 from https://willmaster.com/library/web-development/linking-without-an-a-tag.php.

[72] SimilarWeb LTD. 2020. Top websites in united states. Retrieved 11/28/2020 from https://www.similarweb.com/top-websites/united-states/.

[73] Christopher D. Manning, Prabhakar Raghavan, and Hinrich Schütze. 2008. *Introduction to Information Retrieval*. Cambridge University Press, New York, NY, USA.

[74] C. Matte, N. Bielova, and C. Santos. 2020. Do cookie banners respect my choice? measuring legal compliance of banners from IAB europe's transparency and consent framework. In *2020 IEEE Symposium on Security and Privacy*, 791–809. DOI: 10.1109/SP40000.2020.00076.

[75] Hassan Metwalley, Stefano Traverso, and Marco Mellia. 2015. Unsupervised detection of web trackers. In *2015 IEEE Global Communications Conference*, 1–6. DOI: 10.1109/GLOCOM.2015.7417499.

[76] Microsoft. 2020. Microsoft/playwright-python. Retrieved 12/18/2020 from https://github.com/microsoft/playwright-python.

[77] Microsoft. 2021. Tracking prevention. Retrieved 06/26/2021 from https://docs.microsoft.com/en-us/microsoft-edge/web-platform/tracking-prevention.

[78] Moz Inc. 2021. URL structure. Moz. Retrieved 07/09/2021 from https://moz.com/learn/seo/url.

[79] Mozilla. 2021. Enhanced tracking protection in firefox for desktop. Retrieved 06/26/2021 from https://support.mozilla.org/en-US/kb/enhanced-tracking-protection-firefox-desktop#w_what-enhanced-tracking-protection-blocks.

[80] Mozilla. 2021. Node.textContent - web APIs | MDN. Retrieved 07/09/2021 from https://developer.mozilla.org/en-US/docs/Web/API/Node/textContent.

[81] Kanthashree Mysore Sathyendra, Shomir Wilson, Florian Schaub, Sebastian Zimmeck, and Norman Sadeh. 2017. Identifying the provision of choices in privacy policy text. In EMNLP. Association for Computational Linguistics, 2774–2779.

[82] Niclas. 2020. Clefspeare13/pornhosts. Retrieved 12/13/2020 from https://github.com/Clefspeare13/pornhosts.

[83] Panagiotis Papadopoulos, Nicolas Kourtellis, and Evangelos Markatos. 2019. Cookie synchronization: everything you always wanted to know but were afraid to ask. In *The World Wide Web Conference*. ACM, New York, NY, USA, 1432–1442. DOI: 10.1145/3308558.3313542.

[84] Paul E. Black. 2004. Ratcliff/obershelp pattern recognition. Retrieved 01/02/2021 from https://xlinux.nist.gov/dads/HTML/ratcliffObershelp.html.

[85] Python-Markdown. 2021. Markdown. Retrieved 07/01/2021 from https://github.com/Python-Markdown/markdown.

[86] Ralph Weischedel, Martha Palmer, Mitchell Marcus, Eduard Hovy, Sameer Pradhan, Lance Ramshaw, Nianwen Xue, Ann Taylor, Jeff Kaufman, Michelle Franchini, Mohammed El-Bachouti, Robert Belvin, and Ann Houston. 2013. OntoNotes Release 5.0 LDC2013T19. (2013).

[87] Franziska Roesner, Tadayoshi Kohno, and David Wetherall. 2012. Detecting and Defending Against Third-Party Tracking on the Web. In 155–168.

[88] T. Sakamoto and M. Matsunaga. 2019. After GDPR, Still Tracking or Not? Understanding Opt-Out States for Online Behavioral Advertising. In *2019 IEEE Security and Privacy Workshops (SPW)*, 92–99. DOI: 10.1109/SPW.2019.00027.

[89] Iskander Sanchez-Rola, Matteo Dell'Amico, Platon Kotzias, Davide Balzarotti, Leyla Bilge, Pierre-Antoine Vervier, and Igor Santos. 2019. Can i opt out yet? GDPR and the global illusion of cookie control. In *Proceedings of the 2019 ACM Asia Conference on Computer and Communications Security*. ACM, New York, NY, USA, 340–351. DOI: 10.1145/3321705.3329806.

[90] Alireza Savand. 2021. Html2text. Retrieved 07/01/2021 from https://github.com/Alir3z4/html2text.

[91] SEOPressor. 2019. Does URL structure affect SEO? Retrieved 07/09/2021 from http://seopressor.com/blog/url-structure-affect-seo/.

[92] Peng Shi and Jimmy Lin. 2019. Simple BERT Models for Relation Extraction and Semantic Role Labeling. *arXiv:1904.05255 [cs]*.

[93] Anastasia Shuba and Athina Markopoulou. 2020. NoMoATS: Towards Automatic Detection of Mobile Tracking. *Proceedings on Privacy Enhancing Technologies*, 2020, 2, 45–66. DOI: 10.2478/popets-2020-0017.

[94] Yannis Smaragdakis, Jacob Evans, Caitlin Sadowski, Jaeheon Yi, and Cormac Flanagan. 2012. Sound predictive race detection in polynomial time. In *Proceedings of the 39th annual ACM SIGPLAN-SIGACT symposium on Principles of programming languages*. ACM, New York, NY, USA, 387–400. DOI: 10.1145/2103656.2103702.

[95] Statista. 2022. Internet users in the world 2022. Statista. Retrieved 04/28/2022 from https://statista.com/statistics/617136/digital-population-worldwide/.

[96] Taboola. 2022. Taboola Access Request. Retrieved 08/02/2022 from https://web.archive.org/web/20220710192140/https://accessrequest.taboola.com/access.

[97] Tobias Urban, Martin Degeling, Thorsten Holz, and Norbert Pohlmann. 2020. Beyond the front page:measuring third party dynamics in the field. In *Proceedings of The Web Conference 2020*. ACM, New York, NY, USA, 1275–1286. DOI: 10.1145/3366423.3380203.

[98] Pelayo Vallina, Alvaro Feal, Julien Gamba, Narseo Vallina-Rodriguez, and Antonio Fernández Anta. 2019. Tales from the porn: a comprehensive privacy analysis of the web porn ecosystem. In *Proceedings of the Internet Measurement Conference*. ACM, New York, NY, USA, 245–258. DOI: 10.1145/3355369.3355583.

[99] Zhiju Yang and Chuan Yue. 2020. A comparative measurement study of web tracking on mobile and desktop environments. *Proceedings on Privacy Enhancing Technologies*, 2020, 2, 24–44. DOI: 10.2478/popets-2020-0016.