Letters

RESEARCH LETTER

Prevalence of Third-Party Tracking on COVID-19–Related Web Pages

The internet provides ready access to information related to coronavirus disease 2019 (COVID-19). With a simple web search, individuals can find symptom checkers, locate testing sites, and get tips for keeping themselves safe.

However, online information seeking related to COVID-19 may carry privacy risks. Prior research has shown that web pages visited by individuals seeking health information frequently contain code that initiates data transfers to third parties, such as online advertisers.[1] These transfers often include URLs of visited pages and users' IP addresses. When third parties have code on multiple web pages, they can build detailed profiles of specific individuals' browsing behaviors and interests. This practice, known as "web tracking," can reveal sensitive information about individuals' health conditions and concerns to parties who wish to profit from it.[1]

To better understand the privacy risks of online information seeking related to COVID-19, we assessed the prevalence and characteristics of web tracking on COVID-19–related web pages.

Methods | To identify web pages likely to be visited by individuals seeking COVID-19–related information, we used Google Trends to identify the top 25 search queries related to *COVID* and *coronavirus* in the US on May 15, 2020. We retrieved the top 20 URLs for each query using nonpersonalized Google searches.

We visited each unique web page using webXray, an automated tool that detects third-party tracking on websites.[1] For each web page, we recorded data requests from third-party domains—that is, domains other than that of the website being visited. These requests are significant because they initiate data transfers from a user's computer to third parties. We also recorded the presence of third-party cookies, data stored on a user's computer, which often serve as persistent identifiers that allow users to be tracked across multiple websites.

We calculated the percentages and 95% confidence intervals of web pages that included any third-party data request or any third-party cookie and the median number of third-party data requests and third-party cookies per page, overall and by website type (categorized by top-level domain). We compared results across website types. Using webXray's database of corporate owners of third-party domains, we calculated the most prevalent tracking entities. Analysis was conducted in R version 4.0.2 (R Foundation).

Results | Overall, 535 of 538 (99%; 95% CI, 98%-100%) unique web pages included a third-party data request, with no significant differences by website type, while 477 (89%; 95% CI, 86%-91%) included a third-party cookie (**Table 1**). Compared with commercial web pages, third-party cookies were slightly less common, although still highly prevalent, among government and academic web pages. However, the median numbers of third-party data requests and third-party cookies per page were both higher on commercial web pages (77 requests; 130 cookies) than on government (8 requests; 4 cookies), nonprofit (16 requests; 7 cookies), or academic (14 requests; 10 cookies) web pages.

Most (95%; 95% CI, 93%-97%) web pages included a data request from a third-party domain owned by Google, while 7 other companies received data from at least 40% of web pages studied (**Table 2**).

Discussion | This study found that 99% of COVID-19–related web pages included a third-party data request, and 89% included a third-party cookie. By comparison, a prior study of 1 000 000 popular web pages found that 91% included a third-party data request and 70% included a third-party cookie.[2]

Third-party tracking was pervasive even among government and academic COVID-19–related web pages, on which visitors might reasonably expect greater privacy protections. Decision-makers at these institutions may be unaware of third-party tracking on their websites because they do not realize

Table 1. Third-Party Tracking Overall and by Website Type

	Overall	Commercial (.com, .info, .net)[a]	Government (.gov, .int)[b]	Nonprofit (.org)	Academic (.edu)
No. (%) of web pages	538	320 (59)	110 (20)	91 (17)	17 (3)
Web pages, No. (%) [95% CI]					
With a third-party data request	535 (99) [98-100]	317 (99) [97-100]	110 (100) [96-100]	91 (100) [95-100]	17 (100) [77-100]
With a third-party cookie	477 (89) [86-91]	302 (94) [91-97]	86 (78) [69-85]	78 (86) [76-92]	11 (65) [39-85]
Third-party data requests per page, median (IQR)	30 (1-155)	77 (40-116)	8 (5-13)	16 (7-22)	14 (6-20)
Third-party cookies per page, median (IQR)	29 (0-375)	130 (52-241)	4 (1-4)	7 (2-15)	10 (0-17)

Abbreviation: IQR, interquartile range.

[a] No. of commercial web pages by top-level domain: .com, 311 (97%); .info, 8 (3%); .net, 1 (<1%).

[b] No. of government web pages by level of government: federal, 42 (38%); state, 56 (51%); local, 5 (5%); international, 7 (6%).

Table 2. Most Prevalent Tracking Entities Overall and by Website Type

	Web pages reporting data to a given tracking entity, No. (%) [95% CI]				
Entity	Overall (n = 538)	Commercial (n = 320)	Government (n = 110)	Nonprofit (n = 91)	Academic (n = 17)
Google	507 (95) [93-97]	303 (96) [93-98]	109 (100) [96-100]	78 (86) [76-92]	17 (100) [77-100]
Adobe Systems	267 (50) [46-55]	208 (66) [60-71]	30 (28) [20-37]	20 (22) [14-32]	9 (53) [29-76]
Amazon[a]	260 (49) [45-53]	232 (74) [68-78]	4 (4) [1-10]	19 (21) [13-31]	5 (29) [11-56]
Comscore[b]	249 (47) [43-51]	224 (71) [66-76]	0 (0) [0-4]	24 (26) [18-37]	1 (6) [0-31]
Oracle	246 (46) [42-51]	196 (62) [57-68]	13 (12) [7-20]	32 (35) [26-46]	5 (29) [11-56]
Facebook	243 (46) [41-50]	174 (55) [50-61]	13 (12) [7-20]	48 (53) [42-63]	8 (47) [24-71]
AT&T	227 (43) [38-47]	208 (66) [60-71]	0 (0) [0-4]	13 (14) [8-24]	6 (35) [15-61]
The Trade Desk[c]	212 (40) [36-44]	198 (63) [57-68]	0 (0) [0-4]	8 (9) [4-17]	6 (35) [15-61]
LiveRamp (formerly Acxiom)[d]	204 (38) [34-43]	188 (60) [54-65]	0 (0) [0-4]	9 (10) [5-18]	7 (41) [19-67]
Verizon	187 (35) [31-39]	178 (57) [51-62]	2 (2) [0-7]	3 (3) [1-10]	4 (24) [8-50]

[a] Some Amazon domains belong to the Web Services division and may be for first-party content hosted in an Amazon data center, but the majority of requests were for Amazon-owned marketing services.

[b] Comscore is a cross-platform measurement firm that "provides independent data, metrics, products and services to clients in the media, advertising and marketing industries."[4]

[c] The Trade Desk is a technology firm that "provides a self-service platform that enables clients to purchase and manage digital advertising campaigns across various advertising formats."[5]

[d] LiveRamp is a software and a service company that supports "people-based marketing initiatives across digital channels."[6]

that tools used to monitor website traffic transmit data to third parties.

This study had limitations. First, only 2 mechanisms of third-party tracking were investigated. Because other means of third-party tracking exist, including some designed to evade automated capture, these findings likely underestimate the extent of third-party tracking. Second, because this study was limited to web pages that appeared in the top 20 results for a given Google query, findings may not generalize to web pages with lower search rankings or searches performed using other search engines.

Amid debate and legislative activity focused on the privacy implications of COVID-19 contact-tracing apps, these findings suggest that attention should also be paid to privacy risks of online information seeking.[3]

Matthew S. McCoy, PhD
Timothy Libert, PhD
David Buckler, MUSA
David T. Grande, MD, MPA
Ari B. Friedman, MD, PhD

Author Affiliations: Perelman School of Medicine, University of Pennsylvania, Philadelphia (McCoy, Buckler, Grande, Friedman); School of Computer Science, Carnegie Mellon University, Pittsburgh, Pennsylvania (Libert).

Corresponding Author: Matthew S. McCoy, PhD, Department of Medical Ethics & Health Policy, Perelman School of Medicine, University of Pennsylvania, 423 Guardian Dr, Blockley Hall, Philadelphia, PA 19104 (mmcco@pennmedicine.upenn.edu).

Accepted for Publication: August 10, 2020.

Published Online: September 8, 2020. doi:10.1001/jama.2020.16178

Author Contributions: Drs McCoy and Friedman had full access to all of the data in the study and take responsibility for the integrity of the data and the accuracy of the data analysis.
Concept and design: McCoy, Grande, Friedman.
Acquisition, analysis, or interpretation of data: All authors.
Drafting of the manuscript: McCoy.
Critical revision of the manuscript for important intellectual content: All authors.
Statistical analysis: Libert, Buckler, Friedman.
Administrative, technical, or material support: McCoy, Libert, Grande.
Supervision: McCoy, Friedman.

Conflict of Interest Disclosures: Dr McCoy reported being an uncompensated member of the University of Pennsylvania's Data Ethics Working Group, funded in part through industry gifts to the university. Dr Libert reported receipt of grants from the Defense Advanced Research Projects Agency, CyLab Security and Privacy Institute, and Carnegie Mellon University and consulting with litigants and regulators on matters related to online privacy. No other disclosures were reported.

1. Libert T. Privacy implications of health information seeking on the web. *Commun ACM*. 2015;58(3):68-77. doi:10.1145/2658983

2. Libert T. An automated approach to auditing disclosure of third-party data collection in website privacy policies. In: *Proceedings of the 2018 World Wide Web Conference*. International World Wide Web Conferences Steering Committee; 2018:207-216. doi:10.1145/3178876.3186087

3. US Senate Committee on Commerce, Science, and Transportation. Committee leaders introduce data privacy bill. Published May 7, 2020. Accessed July 2, 2020. https://www.commerce.senate.gov/2020/5/committee-leaders-introduce-data-privacy-bill

4. Reuters Knowledge Direct. Comscore Inc. July 27, 2020. Accessed July 28, 2020. https://advance-lexis-com.proxy.library.upenn.edu/api/document?collection=company-financial&id=urn:contentItem:5WMJ-T391-JBT0-V1PC-00000-00&context=1516831

5. Reuters Knowledge Direct. Trade Desk Inc. July 27, 2020. Accessed July 28, 2020. https://advance-lexis-com.proxy.library.upenn.edu/api/document?collection=company-financial&id=urn:contentItem:5WMJ-Y7P1-JBT0-V3H7-00000-00&context=1516831

6. Reuters Knowledge Direct. LiveRamp Holdings Inc. February 24, 2020. Accessed July 29, 2020. https://advance-lexis-com.proxy.library.upenn.edu/api/document?collection=company-financial&id=urn:contentItem:5WMJ-NDY1-DY82-J10S-00000-00&context=1516831