JAMIA Open, 4(4), 2021, 1–6
https://doi.org/10.1093/jamiaopen/ooab100
Advance Access Publication Date: 9 November 2021
Research and Applications

Research and Applications

Dark clouds and silver linings: impact of COVID-19 on internet users' privacy

Ram D. Gopal[1], Hooman Hidaji[2], Raymond A. Patterson[2], and Niam Yaraghi [3,4]

[1]Information Systems and Management, Warwick Business School, University of Warwick, Coventry, UK, [2]Business Technology Management, Haskayne School of Business, University of Calgary, Calgary, Canada, [3]Business Technology, Miami Herbert Business School, University of Miami, Coral Gables, Florida, USA, and [4]Center for Technology Innovation, Governance Studies, The Brookings Institution, Washington, District of Columbia, USA

Corresponding Author: Niam Yaraghi, PhD, 5250 University Drive, Coral Gables, FL 33146, USA; niamyaraghi@miami.edu

Received 7 September 2021; Revised 30 October 2021; Editorial Decision 2 November 2021; Accepted 5 November 2021

ABSTRACT

Objectives: To examine the impact of coronavirus disease 2019 (COVID-19) pandemic on the extent of potential violations of Internet users' privacy.
Materials and Methods: We conducted a longitudinal study of the data sharing practices of the top 1000 websites in the United States between April 9 and August 27, 2020. We fitted a conditional latent growth curve model on the data to examine the longitudinal trajectory of the third-party data sharing over the 21 weeks period of the study and examine how website characteristics affect this trajectory. We denote websites that asked for permission before placing cookies on users' browsers as "privacy-respecting."
Results: As the weekly number of COVID-19 deaths increased by 1000, the average number of third parties increased by 0.26 (95% confidence interval [CI] 0.15–0.37) $P < 0.001$ units in the next week. This effect was more pronounced for websites with higher traffic as they increased their third parties by an additional 0.41 (95% CI 0.18–0.64); $P < 0.001$ units per week. However, privacy respecting websites that experienced a surge in traffic reduced their third parties by 1.01 (95% CI −2.01 to 0); $P = 0.05$ units per week in response to every 1000 COVID-19 deaths in the preceding week.
Discussion: While in general websites shared their users' data with more third parties as COVID-19 progressed in the United States, websites' expected traffic and respect for users' privacy significantly affect such trajectory.
Conclusions: Attention should also be paid to the impact of the pandemic on elevating online privacy threats, and the variation in third-party tracking among different types of websites.

Key words: privacy, COVID-19, third-parties

Lay Summary
As the coronavirus disease 2019 pandemic progressed in the country, the demand for online services surged. As the level of Internet use increased, websites' opportunity to track and monetize users' data increased with it. In this research, we examine the extent to which websites increased the number of third parties with which they share their users' data and how such practices were moderated by a website's level of respect for users' privacy and traffic surge. We find that while the number of third parties increased over time, the websites with higher respect for privacy tend to decrease the number of their parties only if they also experience a significant increase in their traffic.

INTRODUCTION

The spread of coronavirus disease 2019 (COVID-19) has led many individuals to seek online alternatives for many of their offline activities. This happens for 2 reasons. First, the increase in COVID-19 deaths would lead the local governments to adopt stricter lockdown policies so that people would *have to* use online alternatives. For example, the surge in online streaming services was in part due to the fact that movie theaters were shutdown. Second, the news about the rising COVID-19 deaths would heighten people's perception of the threat of the disease, leading them to be more cautions and *choose* online alternatives. For example, while grocery stores were never shutdown, many customers preferred to do their shopping online out of precaution. Both of these factors increase the demand of online alternatives, and subsequently, online traffic. Figure 1 presents the mediated process through which we hypothesize that an increase in COVID-19 deaths leads to an increase in online traffic.

Websites commonly track their users and share their data with third parties,[1] significantly elevating their users' privacy risks.[2] The increase in online traffic and the subsequent potential for privacy risks exacerbate the current privacy concerns about the digital surveillance practices put in place to combat the COVID-19 pandemic.[3,4] McCoy et al[5] recently discovered that third-party tracking was prevalent among COVID-19-related websites. While that study sheds significant light on our understanding of data sharing practices of healthcare-related websites during a single point in time, it did not examine how such practices evolve over time, especially among a broader range of websites. We bridge these gaps by examining a large sample that is inclusive of a wide variety of websites and studying how they change their data sharing practices over time as COVID-19 spreads in the US, and if such changes can be explained by websites' different features.

OBJECTIVE

A surge in the number of users would allow a website to collect more data and therefore increase the revenue it earns from sharing such data with third parties. A surge in the number of visitors could also increase website's revenue from other streams such as subscriptions and product sales. We intend to examine if websites, upon experiencing a surge in user traffic, reduce their third-party data sharing and instead use the increased revenue from other sources (subscriptions, sales, etc.) to substitute their revenue from data sharing, or if they seize the chance to increase their revenue by even more aggressive data sharing practice. We denote websites that asked for permission before placing cookies on users' browsers as "privacy-respecting." We hypothesize that the level of a website's respect for their users' privacy could influence this decision; the websites that have higher levels of respect for their users' privacy tend to reduce their third parties if they experience a surge in the number of their visitors while the ones with less concern for their users' privacy would see the surge in their visitors as an opportunity to increase their revenue through more aggressive data sharing practices.

MATERIALS AND METHODS

We collected data on the number of third-party hypertext transfer protocol (HTTP) requests (hereafter referred to as third parties) for each of the top 1000 websites (based on ranking data from Alexa Internet) in the United States on a daily basis between April 9 and August 27, 2020, from a virtual server with a New York Internet Protocol (IP) address. Libert[6] provides a detailed description of HTTP requests and how they can be used to track users. As he puts it, "a piece of content from an external server may be called a third-party element [. . .], and the process of downloading such an element is the result of a third-party request. Every time an HTTP request is made, information about the user is transmitted to the server hosting the content. These data include the IP address of the computer making the request, the date and time the request was made, and the type of computer and Web browser employed by the user, which is known as the user-agent field. [. . .] If the server has many such records, patterns of behavior may be attributed to the same combination of IP and user-agent information. This is the most basic form of tracking and is common to all HTTP requests made on the Web."

Note that in many cases, the number of HTTP requests is greater than the number of companies that are making such requests. This could happen for different reasons including when a company sends multiple requests or when different link addresses and third parties belong to the same company. As mentioned earlier, in this research we count the number of HTTP requests during each website visits as it is not only analytically easier to simply count all the requests, but also theoretically, the number of requests is a good measure of the extent of "potential privacy violation, regardless of the number of unique companies that make such requests." We manually tagged the industry that each website operates in and whether it asks for permission before placing cookies on the users' browsers.

To systematically develop a system for classifying websites' industries, all the coauthors reviewed the first 100 websites and independently assigned them to an industry. We then compared our classifications with each other and discussed the websites for which our classifications of industries were different. Once we reached a consensus, we hired a research assistant to label the rest of the web-

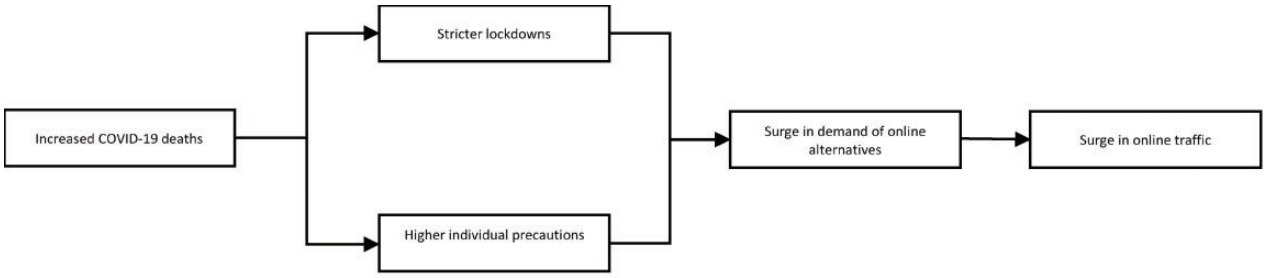

Figure 1. The mediated process through which an increase in COVID-19 deaths leads to an increase in online traffic.

sites. One of the co-authors supervised the research assistant and helped her with classification of the borderline websites.

Feldmann et al[7] show that entertainment categories (which includes streaming and video games) experienced a spike in traffic during the pandemic in Europe. Websites in News and Media, Telecommunication (such as zoom and other video conferencing platforms), and Healthcare industries have also been shown to have experienced a significant spike in traffic.[8,9] We therefore classified websites in *News & Media, Health & Healthcare, Telecommunications*, and *Games* industries as "high-traffic" because they are in sectors that experienced a spike in user traffic due to COVID-19 lockdown policies.

We denote websites that asked for permission before placing cookies on users' browsers as "privacy-respecting." We merged these data with those on weekly number of COVID-19-related deaths in the state of New York over the same period. We used the number of deaths in New York to focus on the local effect of the COVID-19 lockdown policies on the re-action of the websites within the same market. Within the country, the pandemic spread in waves. While New York was badly hit with COVID-19, southern states did not experience a significant rise in cases for a couple of months. Therefore, the lockdown policies, and the subsequent surge in traffic was very location specific. For the same reason, we posit that the website behavior is also dependent on the location of the user; they are going to have a different data gathering and sharing policy with users from New York than their Floridian counterparts simply due to the vast differences in the traffic surge patterns. Since our virtual servers employ a New York IP address, we used the number of COVID-19 deaths in New York to have a consistency between the independent and outcome variables.

Note that the data collection process is very resource intensive as we need to visit each of the 1000 websites twice a day (from New York and California servers) for the 110-day duration of the study. That results in 220 000 website-day initial observations. We therefore developed our own script to collect such data using Golang programming language and ran it on Google Chrome browser through chromedp. The cache was automatically cleared after every single website visit in order to remove any potential effect of browsing history on the number and type of third parties. Postgresql and Clickhouse database management systems were used for storing primary data (the list of the top 1000 websites) and recording and analytical processing of the subsequent data on third parties. We rented 2 virtual servers with 16 CPUs, 64 GB of memory, and 1280 GB of storage.

We used a conditional latent growth curve model to examine the longitudinal trajectory of the third-party data sharing over the period of the study using CALIS procedure in SAS. The model allows us to examine how the spread of COVID-19 pandemic affects the third-party tracking and the extent to which it varies by website characteristics.

RESULTS

Table 1 presents the estimation results of 5 different models. The models gradually develop. In the first model, the effect of COVID-19 deaths is considered to be uniform across all types of websites and therefore there are no estimates for the influence of "high-traffic" and "privacy-respecting" variables or their interactions on the effect of COVID-19 deaths. We sequentially expand the models by including the effects of these variables one by one in the subsequent models. The models in Panels 2 and 3 only include the effects of "high-traffic" and "privacy-respecting," respectively, while the model in panel 4 includes the estimates for the effects of both variables together. Our final model is presented in panel 5 and includes the moderation effects of "privacy-respecting" and "high-traffic" status of websites and their interaction on the association between COVID-19 weekly deaths and the average number of third parties in the subsequent week. Such gradual development of the model allows us to compare the estimates of the same variables across different models and examine the robustness of our model to inclusion of different variables.

As shown in the fifth panel, latent slope is positive, showing that on average, websites increased their third parties by 0.18 (95% confidence interval [CI] 0.11–0.25); $P < 0.001$ units per week. The websites that were expected to receive a higher traffic had a significantly larger intercept (30.08 [95% CI 25.13–35.03]; $P < 0.001$) indicating that they shared their users' data with more third parties at the beginning of the study. Those websites continued to increase their third parties during the study by 0.41 (95% CI 0.26–0.56); $P < 0.001$ units per week. As the weekly number of COVID-19 deaths increased by 1,000, the average number of third parties increased by 0.26 (95% CI 0.15–0.37) $P < 0.001$ units in the next week. This effect was more pronounced for websites with higher traffic as they increased their third parties by an additional 0.41 (95% CI 0.18–0.64); $P < 0.001$ units per week.

Merely asking for consent may not necessarily reduce the number of third parties. In Table 1, the coefficient of "Privacy-respecting vs. other" is not significant in any of the models, meaning that compared with others, the websites that ask for consent do not have a lower number of third parties. However, the moderating effect of "High-traffic & Privacy-respecting vs. others" is significant and negative, meaning that asking for consent is indicative of lower third parties only for those websites that experienced a surge in their traffic. Specifically, if websites that asked for consent also experienced higher traffic, then they would reduce their third parties by 1.22 (95% CI −1.88 to −0.56); $P < 0.001$ units per week. They reduced their third parties by another 1.01 (95% CI −2.01 to 0); $P = 0.05$ units per week in response to every 1000 COVID-19 deaths in the preceding week.

DISCUSSION

One limitation of this analysis is the fact that all websites within the same industry are assumed to have experienced the same level of change in their traffic. This is a crude measure as it does not take into account the variations in traffic shifts of different websites. A better alternative would have been to individually determine traffic trends for each website.

The other limitation arises from the fact that the number of third parties of a single website would fluctuate over time. We analyze the daily average number of third parties across the 1000 websites. This would decrease the effect of potential fluctuations. Moreover, we track the same websites over a long period of time. Despite the temporal changes, the number of third parties of a single website tend to fluctuate around an average trend. A better solution would have been to measure the number of third parties over multiple times within the day and then take the average as the main outcome variable.

Table 1. Parameter estimates of the latent growth curve in third-party tracking among the top US websites

	Panel 1		Panel 2		Panel 3		Panel 4		Panel 5	
Coefficient	Estimate (95% CI)	*P*-value	Estimate (95% CI)	*P*-value	Estimate (95% CI)	*P*-value	Estimate (95% CI)	*P*-value	Estimate (95% CI)	*P*-value
Latent intercept										
Main effect	20.33 (18.04 to 22.61)	<0.001	20.19 (17.9 to 22.48)	<0.001	20.33 (18.04 to 22.61)	<0.001	20.19 (17.91 to 22.48)	<0.001	20.17 (17.89 to 22.46)	<0.001
High-traffic vs. other	29.34 (24.41 to 34.28)	<0.001	29.98 (25.03 to 34.93)	<0.001	29.34 (24.41 to 34.28)	<0.001	29.98 (25.03 to 34.93)	<0.001	30.08 (25.13 to 35.03)	<0.001
Privacy-respecting vs. other	0.56 (−8.04 to 9.16)	0.898	0.56 (−8.04 to 9.16)	0.898	0.51 (−8.11 to 9.14)	0.907	0.55 (−8.07 to 9.18)	0.9	0.83 (−7.8 to 9.46)	0.85
High-traffic and privacy-respecting vs. other	6.88 (−14.74 to 28.5)	0.533	6.88 (−14.74 to 28.51)	0.533	6.88 (−14.74 to 28.5)	0.533	6.88 (−14.74 to 28.51)	0.533	5.1 (−16.6 to 26.79)	0.645
Latent slope										
Main effect	0.21 (0.14 to 0.27)	<0.001	0.18 (0.11 to 0.25)	<0.001	0.21 (0.14 to 0.27)	<0.001	0.18 (0.11 to 0.25)	<0.001	0.18 (0.11 to 0.25)	<0.001
High-traffic vs. other	0.29 (0.15 to 0.42)	<0.001	0.39 (0.24 to 0.54)	<0.001	0.29 (0.15 to 0.42)	<0.001	0.39 (0.24 to 0.54)	<0.001	0.41 (0.26 to 0.56)	<0.001
Privacy-respecting vs. other	0.13 (−0.11 to 0.36)	0.285	0.13 (−0.11 to 0.36)	0.285	0.12 (−0.14 to 0.38)	0.362	0.13 (−0.13 to 0.39)	0.337	0.17 (−0.09 to 0.44)	0.197
High-traffic and privacy-respecting vs. other	−0.92 (−1.52 to −0.33)	0.002	−0.92 (−1.52 to −0.33)	0.002	−0.92 (−1.52 to −0.33)	0.002	−0.92 (−1.52 to −0.33)	0.002	−1.22 (−1.88 to −0.56)	<0.001
COVID-19 deaths (in 1000)										
Main effect	0.35 (0.25 to 0.44)	<0.001	0.27 (0.17 to 0.37)	<0.001	0.35 (0.25 to 0.44)	<0.001	0.27 (0.17 to 0.38)	<0.001	0.26 (0.15 to 0.37)	<0.001
High-traffic vs. other			0.36 (0.14 to 0.58)	0.002			0.36 (0.14 to 0.58)	0.002	0.41 (0.18 to 0.64)	<0.001
Privacy-respecting vs. other					−0.03 (−0.4 to 0.34)	0.886	−0.01 (−0.37 to 0.36)	0.974	0.15 (−0.25 to 0.55)	0.452
High-traffic and privacy-respecting vs. other									−1.01 (−2.01 to 0)	0.05
Variance/covariance										
Latent intercept	855.94 (773.5 to 938.39)	<0.001	855.87 (773.44 to 938.31)	<0.001	855.94 (773.5 to 938.39)	<0.001	855.87 (773.44 to 938.31)	<0.001	855.85 (773.42 to 938.28)	<0.001
Latent slope	0.68 (0.6 to 0.76)	<0.001	0.68 (0.6 to 0.76)	<0.001	0.68 (0.6 to 0.76)	<0.001	0.68 (0.6 to 0.76)	<0.001	0.68 (0.6 to 0.75)	<0.001
Intercept and slope	1.75 (−0.03 to 3.53)	0.054	1.74 (−0.04 to 3.52)	0.055	1.75 (−0.03 to 3.53)	0.054	1.74 (−0.04 to 3.52)	0.055	1.73 (−0.04 to 3.51)	0.056
Residuals										
e1–e21	55.54 (54.32 to 56.76)	<0.001	55.54 (54.32 to 56.76)	<0.001	55.54 (54.32 to 56.76)	<0.001	55.54 (54.32 to 56.76)	<0.001	55.54 (54.32 to 56.76)	<0.001

(continued)

Table 1. continued

Coefficient	Panel 1		Panel 2		Panel 3		Panel 4		Panel 5	
	Estimate (95% CI)	*P*-value	Estimate (95% CI)	*P*-value	Estimate (95% CI)	*P*-value	Estimate (95% CI)	*P*-value	Estimate (95% CI)	*P*-value
FIT summary										
Chi-square	18 686.80		18 610.16		18 620.07		18 610.15		18 606.30	
Chi-square DF	301		298		298		297		296	
Pr > Chi-square	<0.001		<0.001		<0.001		<0.001		<0.001	
Standardized RMR (SRMR)	0.11		0.11		0.10		0.10		0.10	
Goodness of fit index (GFI)	0.34		0.34		0.34		0.34		0.34	
RMSEA estimate	0.26		0.26		0.26		0.26		0.26	
Akaike Information Criterion	18 732.80		18 662.16		18 672.07		18 664.15		18 662.31	
Schwarz Bayesian Criterion	18 842.81		18 786.52		18 796.44		18 793.30		18 796.24	
Bentler Comparative Fit Index	0.73		0.73		0.73		0.73		0.73	
Bentler-Bonett non-normed index	0.75		0.75		0.75		0.75		0.75	
Bollen normed index Rho1	0.75		0.75		0.75		0.75		0.75	

Note: The main results with interaction terms are provided in Panel 5 which includes moderation effects of "privacy-respecting" and "high-traffic" status of websites and their interaction on the association between COVID-19 weekly deaths and the average number of third parties in the subsequent week. In Panels 1 to 4, the parameter estimates are consistent in their sign and significance across various configurations of the model, indicating its robustness.

We removed 117 websites from the sample because either they were adult websites, or we could not collect data on their third parties for the period of the study.

CONCLUSION

Amid national discussions about the potential legislation aimed to protect users' privacy, [10] the insights of our research suggest that attention should also be paid to the impact of the pandemic on elevating online privacy threats, and the variation in third-party tracking among different types of websites.

FUNDING

This work was supported by Social Sciences and Humanities Research Council of Canada (430-2018-00433).

AUTHOR CONTRIBUTIONS

NY drafted the study design, performed the analysis, and drafted the manuscript. RP and HH participated in the study design, supervised data collection, reviewed the analysis, and revised the manuscript. RG participated in the study design, reviewed the analysis, and revised the manuscript. All authors read and approved the final manuscript.

CONFLICT OF INTEREST STATEMENT

The authors have no competing interests to declare.

DATA AVAILABILITY

The data will be shared on reasonable request to the corresponding author.

REFERENCES

1. 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*. 2018:207–216; Lyon, France.
2. Englehardt S, Reisman D, Eubank C, *et al.* Cookies that give you away: the surveillance implications of web tracking. In: *Proceedings of the 24th International Conference on World Wide Web*. WWW '15. International World Wide Web Conferences Steering Committee; 2015:289–299; Florence, Italy. doi:10.1145/2736277.2741679.
3. Bengio Y, Janda R, Yu YW, *et al.* The need for privacy with public digital contact tracing during the COVID-19 pandemic. *Lancet Digit Health* 2020; 2 (7): e342–e344.
4. Sharma T, Bashir M, Bashir M. Use of apps in the COVID-19 response and the loss of privacy protection. *Nat Med* 2020; 26 (8): 1165–7.
5. McCoy MS, Libert T, Buckler D, Grande DT, Friedman AB. Prevalence of third-party tracking on COVID-19-related web pages. *JAMA* 2020; 324 (14): 1462–4.
6. Libert T. Exposing the hidden web: An analysis of third-party HTTP requests on 1 million websites. *arXiv preprint arXiv:1511.00619*. Published online 2015.
7. Feldmann A, Gasser O, Lichtblau F, *et al.* Implications of the COVID-19 Pandemic on the Internet Traffic. In: *Broadband Coverage in Germany; 15th ITG-Symposium*. VDE; 2021:1–5.
8. Dahiya S, Rokanas LN, Singh S, Yang M, Peha JM. Lessons from internet use and performance during COVID-19. *J Inf Policy* 2021; 11: 202–21.
9. Comscore Sees Shifting Category Trends for Digital Audiences Amid Pandemic. Comscore, Inc. https://www.comscore.com/Insights/Blog/Comscore-Sees-Shifting-Category-Trends-for-Digital-Audiences-Amid-Pandemic Accessed October 15, 2021.
10. Committee Leaders Introduce Data Privacy Bill. U.S. Senate Committee on Commerce, Science, & Transportation. Published May 7, 2020. https://www.commerce.senate.gov/2020/5/committee-leaders-introduce-data-privacy-bill Accessed April 15, 2021.