DIW BERLIN

1998

Discussion Papers

Deutsches Institut für Wirtschaftsforschung 2022

Facebook Shadow Profiles

Luis Aguiar, Christian Peukert, Maximilian Schäfer and Hannes Ullrich

IMPRESSUM

DIW Berlin
German Institute for Economic Research
Mohrenstr. 58
10117 Berlin

Tel. +49 (30) 897 89-0
Fax +49 (30) 897 89-200
https://www.diw.de

ISSN electronic edition 1619-4535

Papers can be downloaded free of charge from the DIW Berlin website:
https://www.diw.de/discussionpapers

Discussion Papers of DIW Berlin are indexed in RePEc and SSRN:
https://ideas.repec.org/s/diw/diwwpp.html
https://www.ssrn.com/link/DIW-Berlin-German-Inst-Econ-Res.html

Facebook Shadow Profiles*

Luis Aguiar[a], Christian Peukert[b], Maximilian Schäfer[c], and Hannes Ullrich[d,e]

[a]Department of Business Administration, University of Zurich
[b]Department of Strategy, Globalization and Society, University of Lausanne
[c]Tobin Center for Economic Policy, Yale University
[d]Department of Firms and Markets, DIW Berlin
[e]Department of Economics, University of Copenhagen

February 2022

Abstract

Data is often at the core of digital products and services, especially when related to online advertising. This has made data protection and privacy a major policy concern. When surfing the web, consumers leave digital traces that can be used to build user profiles and infer preferences. We quantify the extent to which Facebook can track web behavior outside of their own platform. The network of engagement buttons, placed on third-party websites, lets Facebook follow users as they browse the web. Tracking users outside its core platform enables Facebook to build shadow profiles. For a representative sample of US internet users, 52 percent of websites visited, accounting for 40 percent of browsing time, employ Facebook's tracking technology. Small differences between Facebook users and non-users are largely explained by differing user activity. The extent of shadow profiling Facebook may engage in is similar on privacy-sensitive domains and across user demographics, documenting the possibility for indiscriminate tracking.

JEL codes: D18, L4, L5, L86
Keywords: facebook, privacy, user data, web tracking, shadow profiles

*We thank Avi Goldfarb, Dirk Bergemann, Joel Waldfogel, Ken Wilbur, Lisa George, Tomaso Duso, and participants of presentations at the Media Economics Workshop Stellenbosch, ETH Zurich Center for Law and Economics, University of Zurich, HEC Lausanne, and IPTS Seville for valuable feedback. We are grateful to Ilia Azizi for excellent research assistance. Peukert acknowledges funding from the Swiss National Science Foundation Grant Number 100013_197807. E-mail: luis.aguiar@business.uzh.ch (Aguiar), christian.peukert@unil.ch (Peukert), maximilian.schaefer@yale.edu (Schäfer), hullrich@diw.de (Ullrich).

1 Introduction

The fundamental business model of many online platforms such as Facebook consists in generating revenue through online advertising. Because detailed information about consumers' types and preferences is crucial for targeted advertising effectiveness, online platforms have developed innovative technologies to collect and analyze behavioral data (Trusov, Ma & Jamal 2016).

Facebook's engagement buttons are an important example. Using cookies, Facebook can track users across all websites on which a Facebook "Like" or "Share" button appears, even if users never actively click on them (Roosendaal 2012). This, in turn, allows Facebook to build *shadow profiles* (Garcia 2017), regardless of whether an individual ever signed up for a Facebook account. *Shadow profiles* allow Facebook to connect individual browsing behavior to existing and future Facebook accounts.[1] They can also be used to place targeted advertisements through Facebook's cross-site advertising network, or to predict missing data points on Facebook users with similar browsing characteristics. In fact, empirical research has shown that even for platforms with extensive internal usage-based information – such as search engines and social networking websites – access to data from outside their core platform helps improve their predictions of user profiles (Trusov, Ma & Jamal 2016).

A large number of websites have integrated Facebook's engagement buttons since their launch in 2009. Websites implement Facebook buttons perhaps with the intent to increase traffic through social media referrals (Sismeiro & Mahmood 2018). Theoretical work also suggests that the widespread adoption of engagement buttons can be the result of a prisoner's dilemma (Krämer, Schnurr & Wohlfarth 2019). The widespread diffusion of Facebook's engagement buttons and tracking scripts on the web has been largely documented (Chaabane, Kaafar & Boreli 2012, Libert 2015, Lerner et al. 2016). In 2020, Facebook's engagement buttons had already reached 40 percent of the top 1'000 websites (Australian Competition and Consumer Commission 2021). Existing studies, however, typically focus on the share of websites being monitored by such tracking technologies, leaving aside the share of individual users' activity that can be tracked by online platforms such as Facebook.

The extent of shadow profiling by Facebook has raised substantial public interest.[2] In 2018, when questioned in U.S. congress and European parliament hearings, CEO Mark Zuckerberg replied that he did not know about shadow profiles nor how much data on

[1] See also Facebook's official press release on data collection outside of Facebook at https://about.fb.com/news/2018/04/data-off-facebook/, accessed 24 September 2021.

[2] See, for example, https://eu.usatoday.com/story/tech/columnist/baig/2018/04/13/how-facebook-can-have-your-data-even-if-youre-not-facebook/512674002/, accessed 24 September 2021.

non-Facebook users was collected.[3] Despite its relevance for both public policy as well as advertisers, the magnitude of users' online attention being tracked by Facebook remains largely undocumented.

Against this backdrop, the goal of this paper is to tackle four main questions. First, what is the share of individuals' browsing activity that can be tracked by Facebook? Second, how do these shares vary between individuals who are users of Facebook and those who are not? In other words, we aim at quantifying Facebook's ability to create *shadow profiles* of individuals who are likely unaware and have not consciously consented to Facebook tracking their online activities. Third, how do these differences vary across demographic groups? Fourth, does Facebook's ability to track user behavior vary across different types of websites?

Our empirical analysis relies on the entire browsing history of about 5'000 representative U.S. internet users in 2016, combined with information on whether the visited websites interacted with Facebook servers through engagement buttons. This allows us to measure Facebook's ability to track users across the web. Specifically, we estimate the share of online activity (measured in visits and time spent online) that can be tracked by Facebook, distinguishing between users and non-users of Facebook. In addition, we document the extent of tracking by user demographics, browsing intensity, and types of domains.

Our results show that Facebook's technology is present on 52% of websites visited, accounting for 40% of browsing time. Facebook can track 55% of the websites visited by Facebook users, and 44% of non-Facebook users, which amounts to 41% and 38% of browsing time, respectively. While websites visited by Facebook users are covered more by Facebook's engagement buttons, this difference largely disappears after controlling for browsing time. The absence of differences in browsing time tracked is robust across users with varying demographics. Facebook's ability to track is prevalent even on privacy-sensitive websites but less so in website categories likely to be competing with Facebook. These findings provide a core insight: irrespective of the decision to avoid Facebook, demographics, or the websites users decide to visit, Facebook can follow close to half of users' attention online.

Our results have important implications for ongoing debates around privacy regulation and competition policy. For individual users, privacy concerns could be reduced if tracking by Facebook outside of its own platform could easily be avoided by quitting Facebook or by visiting websites without engagement buttons. While recent privacy regulations and industry changes have made third party tracking more difficult (Johnson,

[3]See https://venturebeat.com/2018/05/22/mark-zuckerberg-dodges-question-from-european-parliament-on-facebook-shadow-profiles/ and https://techcrunch.com/2018/04/11/facebook-shadow-profiles-hearing-lujan-zuckerberg, accessed 24 September 2021.

Shriver & Goldberg 2020, Peukert et al. 2022), the reality of tracking remains much the same. Legal ascertainment of consumer consent has become more robust over time but consumers effectively still do not know they consented or have little choice.[4] The extent of tracking we document is indicative of the potential privacy risk due to shadow profiling for advertising purposes irrespective of technological or legal barriers.

2 Data and measurement

We have access to individual-level desktop browsing data of a representative sample of the U.S. population via the market research firm Nielsen. Participants are incentivized to install a software that records all web browsing activity and fill in a survey of basic demographics, such as gender, employment, age, education, and income. We classify individuals who refused to answer a question about their income as privacy sensitive (Goldfarb & Tucker 2012). For each user, we observe the web addresses (URLs) of all websites visited in 2016, as well as the time spent on each URL. We define Facebook users as those who visit Facebook at least once in that year. Nielsen further provides a categorization of websites on which we base our analysis of privacy-sensitive websites and competitors to Facebook.

Historical information on websites' connections to Facebook comes from the HTTP-Archive project, which periodically crawls a large number of websites to record data on their use of third-party technology (including cookies). We use information collected on June 1, 2016 and filter connections to Facebook servers.

Combining the two datasets, we can compute the share of websites a user visits that make requests to Facebook servers, i.e. the share that can be tracked by Facebook. Specifically, we jointly observe, for each user, the share of domains and browsing time tracked Y, browsing intensity W, discretized personal characteristics X, Facebook user status D, and visited website categories K. We provide estimates of $E[Y|W = w_p, D]$ and $E[Y|W = \bar{w}, D, X]$, where w_p denotes browsing intensity by percentiles $p \in \{0.2, 0.4, 0.6, 0.8, 1.0\}$ and $\bar{w}$ denotes mean browsing intensity.

One potential concern is that some of the websites visited by individual users in the clickstream data are not observed in the HTTPArchive project. In other words, we are unable to obtain websites' connections to Facebook for a subset of domains in our clickstream data. In order to quantify the extent of this selection problem, we estimate Manski bounds (Manski 1995). More specifically, of the 491'841 website domains visited

[4]The Facebook antitrust case at the German federal cartel office was focused on consent and consumer choice, see https://www.reuters.com/article/us-facebook-germany-idUSKCN1PW0SW, accessed 24 September 2021.

by 4,989 users in our data, 109'512 are observed in HTTPArchive, ($z = 1$). Domains not observed in HTTPArchive, ($z = 0$), account for $P(z = 0) = 0.25$ of unique domains visited and $P(z = 0) = 0.11$ of total browsing time. Decomposing $E[Y] = E[Y|z = 1]P(z = 1) + E[Y|z = 0]P(z = 0)$, we note $E[Y|z = 0]$ cannot be estimated. However, we know that $0 \leq E[Y|z = 0] \leq 1$ and so can estimate bounds of $E[Y]$: $E[Y|z = 1]P(z = 1) \leq E[Y] \leq E[Y|z = 1]P(z = 1) + P(z = 0)$.

A final concern is that internet users may explicitly opt out of web tracking. However, awareness of tracking and opting out is rare in practice and requires significant user sophistication (Melicher et al. 2016, Mathur et al. 2018, Weinshel et al. 2019, Johnson, Shriver & Du 2020).

3 Results

3.1 Shadow profiles across different users

Figures 1 and 2 show the extent to which Facebook can track its users and non-users, and how this depends on overall browsing intensity and demographics.

The top panel in Figure 1 focuses on the total number of websites that individuals visit during the sample period, excluding Facebook. As shown by the two vertical dashed lines, 55% of websites visited by Facebook users and 44% of websites visited by non-Facebook users are tracked, on average, by Facebook. The share of websites visited by an individual that can be tracked by Facebook increases with total online activity, measured in quintiles of the total number of websites visited (including Facebook). The extent of Facebook's potential tracking – conditional on online activity — is similar for users and non-users of Facebook. Hence, differences in the extent of overall tracking, denoted by the dashed vertical lines, are due to Facebook users' higher online activity.

The number of visited websites may not fully capture the intensity with which individuals browse the internet. In the lower panel we report mean shares of total browsing time that Facebook can track over the sample period. We observe smaller differences in the average share of online activity that can be tracked for users and non-users of Facebook, depicted as vertical dashed lines at 41% and 38%. This suggests that Facebook is similarly well connected to websites on which users and non-users of Facebook spend a similar amount of time. Perhaps surprisingly, the share of an individual's online browsing time that can be tracked by Facebook increases only marginally with online activity.

For website domains visited, the estimated Manski bounds are $[0.38, 0.64]$ and for browsing time $[0.35, 0.47]$. Absent any assumption on the selection of domains, these

bounds remain informative and close to our point estimates. We therefore focus on point estimates when reporting all further results.

Figure 2 shows the mean share of websites and browsing time that Facebook can track by demographic groups. The top panel shows some heterogeneity regarding the extent to which certain demographic groups are being tracked. For instance, females tend to be tracked more than male, while we see little difference in tracking between individuals who are privacy sensitive and those who are not. We do not find significant differences in the share of domains tracked between users and non-users of Facebook for any demographic group. As these shares are conditional on mean browsing intensity, any differences would be ascribed to the types of websites visited by users, in terms of Facebook coverage, in their respective demographic group. Hence, Facebook benefits from engagement buttons being placed on a diverse set of websites. The lower panel further shows that, within demographic groups, there are no significant differences in the share of browsing time tracked for users and non-users of Facebook.

3.2 Shadow profiles across types of websites

Table 1 shows the share of websites and browsing time that can be tracked by Facebook, distinguishing between website categories. Looking at columns (1) and (2) we see that the overall share of websites that are connected to Facebook servers is significantly higher for users of Facebook relative to non-users. However, splitting online activity by the type of websites visited reveals that this difference is mainly driven by tracking differences on websites related to instant messaging services, where Facebook users are tracked to a much larger extent than non-Facebook users.[5] For the remainder of the website categories, we observe small differences between users and non-users of Facebook. The shares of websites and browsing time tracked are smaller overall for website categories that likely include competing online platforms. This observation is consistent with a reduced level of integration of Facebook's engagement buttons in the presence of competing websites, in line with individuals' interest in privacy online. When focusing further on browsing time in columns (3) and (4), the differences between users and non-users of Facebook are much less pronounced. This holds overall and in all website categories except instant messaging. Thus, it is apparent that there are differences in Facebook's overall ability to track visits and time spent across website categories. Nevertheless, Facebook's engagement button technology can track attention online irrespective of types of websites, even in categories considered privacy-sensitive. This is true for users and non-users of Facebook alike, since there are little within-categories differences between these two groups.

[5]Note that these instant messaging services do not include services offered by Facebook.

3.3 Shadow profiles in relation to Facebook use

Lastly, we explore the distribution of browsing time across Facebook.com, websites that have Facebook's engagement buttons ("tracked"), and websites that do not have Facebook's engagement buttons ("not tracked"). Figure 3 shows the average amount of time spent on either one of these three categories of website, by demographic. Across most demographic groups, the browsing time that is either directly or indirectly monitored by Facebook accounts for more of 50 percent of the total browsing time.

Across all demographics, the surfing time tracked by Facebook is larger than the actual time spent on Facebook's platform. While we find large variation in total surfing time across groups, the relative shares of surfing time on Facebook and on other websites tracked by Facebook are remarkably similar. Figure 3 highlights the scope of Facebook tracking including the online activity it can directly monitor on its own platform. The network of engagement button appears to increase the amount of time tracked proportionally to the time spent on Facebook.com.

4 Discussion and conclusion

We combine information on websites' implementation of Facebook engagement buttons with detailed individual-level data on the web browsing history of a representative sample of US consumers. This allows us to document the extent to which Facebook can build shadow profiles based on users' web browsing behavior outside of Facebook's core platform. We document an indiscriminate ability to collect user data, independent of user characteristics such as demographics and whether they actively use Facebook's social networking platform.

Our results have a range of implications. Broadly speaking, our results highlight the potential of data to shift economic power. Information that lets firms infer consumer preferences can be used to extract rents, for example through targeted advertising or price discrimination. When information is better available to large firms, this may lead to further concentration in the online advertising industry. Resulting higher prices for advertisers increase their cost which can in turn increase prices for consumers. By documenting that a large platform like Facebook is able to collect data on consumers that do not use their platform, our results suggest further potential harms to consumers. Even users that do not receive utility from using Facebook are subject to the externalities that Facebook's data collection may impose on them.

Regulatory efforts around the globe have tried to reduce the type of de-facto tracking that we highlight above. Examples include the European Union's (EU) ePrivacy Direc-

tive from 2003 that regulated opt-in to data collection via cookies and the EU General Data Protection Regulation from 2018. Among other things, the latter mandated informed consent to data collection and introduced harsh fines. Similar legislation was introduced in the US with the California Consumer Privacy Act that became effective in 2020. Leading actors in the industry such as Apple and Google are recently orchestrating a move away from third-party cookies for tracking. However, the underlying economics of consumer tracking have not changed. Individual data is important for personalized advertising, which is generating enormous revenues for the industry as a whole. Cookies are or will be replaced by new generations of tracking technologies, such as device fingerprinting, software development kits, and in-app browsing. The ability of Facebook to create shadow profiles is therefore not limited to Facebook's current technology of cookies and engagement buttons, and perhaps more importantly not limited to Facebook. Google places third-party scripts on 80% of the top 1000 websites (Australian Competition and Consumer Commission 2021) and has shifted to using its Chrome browser to track online activities. Hence, despite regulatory efforts and technological change, the potential for indiscriminate large-scale tracking, such as using third-party scripts as we document, is likely to remain.

References

Australian Competition and Consumer Commission. 2021. "Digital advertising services inquiry: interim report."

Chaabane, Abdelberi, Mohamed Ali Kaafar, and Roksana Boreli. 2012. "Big friend is watching you: Analyzing online social networks tracking capabilities." *Proceedings of the 2012 ACM Workshop on Online Social Networks*, 7–12.

Garcia, David. 2017. "Leaking privacy and shadow profiles in online social networks." *Science Advances*, 3: e1701172.

Goldfarb, Avi, and Catherine Tucker. 2012. "Shifts in privacy concerns." *American Economic Review: Papers and Proceedings*, 102: 349–353.

Johnson, Garrett A, 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.

Johnson, Garrett, Scott Shriver, and Samuel Goldberg. 2020. "Privacy & market concentration: Intended & unintended consequences of the GDPR."

Krämer, Jan, Daniel Schnurr, and Michael Wohlfarth. 2019. "Winners, losers, and facebook: The role of social logins in the online advertising ecosystem." *Management Science*, 65(4): 1678–1699.

Lerner, Adam, 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." *Proceedings of the 25th USENIX Security Symposium*, 997–1013.

Libert, Timothy. 2015. "Exposing the Hidden Web: An Analysis of Third-Party HTTP Requests on One Million Websites." *International Journal of Communication*, 9: 3544–3561.

Manski, Charles F. 1995. *Identification problems in the social sciences.* Harvard University Press.

Mathur, Arunesh, Jessica Vitak, Arvind Narayanan, and Marshini Chetty. 2018. "Characterizing the use of browser-based blocking extensions to prevent online tracking." *Fourteenth Symposium on Usable Privacy and Security ({SOUPS} 2018)*, 103–116.

Melicher, William, Mahmood Sharif, Joshua Tan, Lujo Bauer, Mihai Christodorescu, and Pedro Giovanni Leon. 2016. "Preferences for Web Tracking." *Proceedings on Privacy Enhancing Technologies*, 2016(2): 1–20.

Peukert, Christian, Stefan Bechtold, Michail Batikas, and Tobias Kretschmer. 2022. "Regulatory Spillovers and Data Governance: Evidence from the GDPR." *Marketing Science*.

Roosendaal, Arnold. 2012. "We Are All Connected to Facebook... by Facebook!" In *European Data Protection: In Good Health?*. , ed. Serge Gutwirth, Ronald Leenes, Paul DeHert and Yves Poullet, 3–19. Springer, New York.

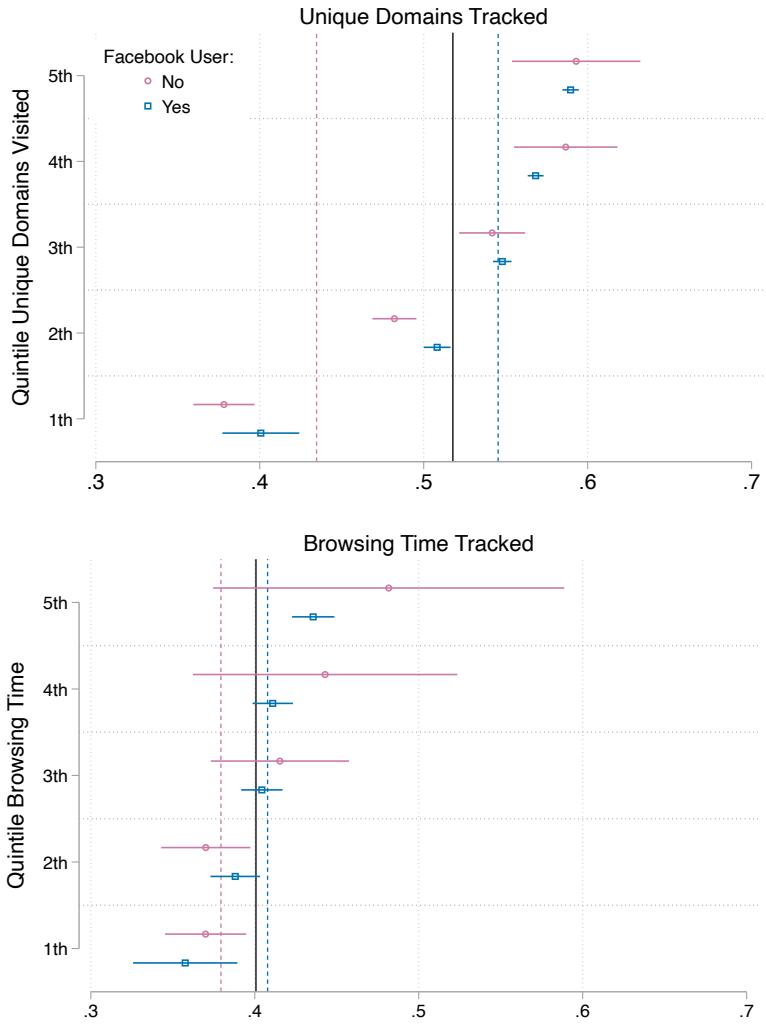
Sismeiro, Catarina, and Ammara Mahmood. 2018. "Competitive vs. complementary effects in online social networks and news consumption: A natural experiment." *Management Science*, 64(11): 5014–5037.

Trusov, Michael, Liye Ma, and Zainab Jamal. 2016. "Crumbs of the cookie: User profiling in customer-base analysis and behavioral targeting." *Marketing Science*, 35(3): 405–426.

Weinshel, Ben, Miranda Wei, Mainack Mondal, Euirim Choi, Shawn Shan, Claire Dolin, Michelle L Mazurek, and Blase Ur. 2019. "Oh, the Places You've Been! User Reactions to Longitudinal Transparency About Third-Party Web Tracking and Inferencing." *Proceedings of the 2019 ACM SIGSAC Conference on Computer and Communications Security*, 149–166.

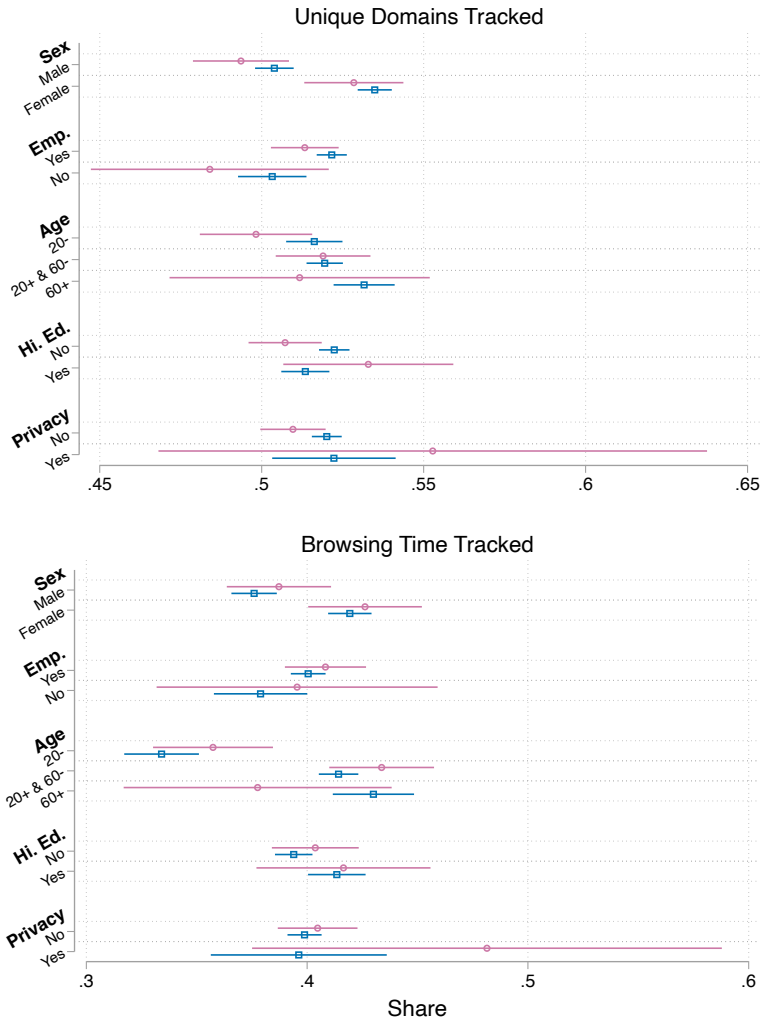
A Figures and Tables

Figure 1: Extent of Facebook web tracking by user type



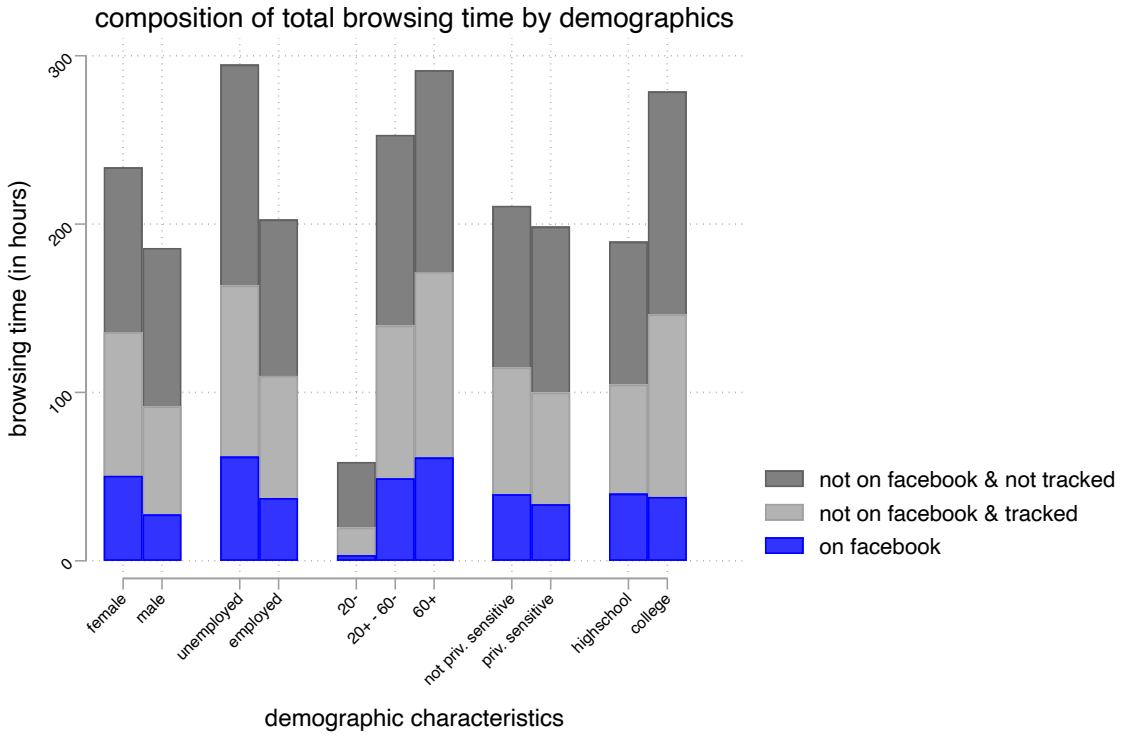
Notes: The Figure shows mean shares of the number of unique visited websites and browsing time tracked by quintiles of each browsing intensity measure, for Facebook users and non-Facebook users. In the first panel, the solid line shows the mean share of websites tracked, 52%, and dashed vertical lines show the mean shares of websites tracked for Facebook users and non-Facebook users, 55% and 44%. In the second panel, the solid line shows the mean share of browsing time tracked, 40%, and dashed vertical lines show the mean shares of browsing time tracked for Facebook users and non-Facebook users, 41% and 38%.

Figure 2: Extent of Facebook web tracking by user type



Notes: The Figure shows mean shares of visited websites and browsing time tracked, conditional on the respective demographic characteristic and mean browsing intensity, for Facebook users and non-Facebook users.

Figure 3: Extent of Facebook web tracking by Facebook usage



Notes: The Figure shows the average amount of browsing time an individual falling into a respective demographic spends on a) Facebook, b) websites with Facebook engagement buttons, and c) websites without Facebook engagement buttons.

Table 1: Shadow profiling by types of websites

	Unique domains		Browsing time	
Website category	non-FB user	FB user	non-FB user	FB user
All*†	0.44	0.55	0.38	0.41
Privacy-sensitive				
Adult*	0.09	0.13	0.09	0.12
Career Development	0.42	0.45	0.37	0.42
Dating	0.54	0.53	0.55	0.55
Finance/Insurance/Investment	0.45	0.46	0.45	0.45
Gambling	0.66	0.62	0.68	0.67
Government	0.15	0.16	0.15	0.15
Health, Fitness & Nutrition	0.63	0.64	0.63	0.64
Real Estate/Apartments	0.74	0.75	0.78	0.79
Competing platforms				
E-mail*	0.01	0.04	0.01	0.01
Instant Messaging*†	0.08	0.23	0.05	0.19
Member Communities	0.58	0.60	0.58	0.58
Search*†	0.21	0.29	0.22	0.29

Notes: The table reports mean user-level shares of unique domains and browsing time tracked by Facebook (FB). $\star$: Difference in unique domains between user groups are statistically significant at the five percent level. †: Difference in browsing time between user groups are statistically significant at the five percent level.