Unaccounted Privacy Violation:

A Comparative Analysis of Persistent Identification of Users Across Social Contexts

By Ido Sivan-Sevilla, Wenyi Chu, Xiaoyu Liang, Helen Nissenbaum

ABSTACT

Cross-contexts inference about individuals is central to the profiling and targeting capacities of the online advertising industry. Advertisers identify moments of vulnerability, in part, by drawing on information gleaned from individuals across social domains. According to the theory of contextual integrity, the failure to respect informational norms governing different social contexts violates people's privacy expectations. Although much valuable work has been written about online tracking and surveillance by commercial actors, this research adds a distinctive perspective. It offers a context-sensitive empirical study of the usage of persistent identifiers by the advertising industry. Previous works studied online tracking in bulk, across thousands of websites, without distinguishing the different social contexts of each website. This study, however, aims to start and build a contextual understanding of tracking practices and highlight differences and similarities in online tracking across social contexts. This is a first demonstration, to the best of our knowledge, of studying online tracking in a contextual perspective. Therefore, we ask: (1) To what extent the social contexts of websites explain the usage of persistent user identifiers by the advertising industry? (2) Should we expect trackers embedded in certain social contexts (e.g. healthcare) to become more likely to link user identities into other social contexts?

To answer these questions, we conducted a context-sensitive empirical analysis of the usage of persistent user identifiers by trackers across social contexts. We used an open-source instrumented FireFox browser (openWPM) in stateful tracking mode to launch six browsing experiments, for all possible sequences of browsing across three social contexts – news, health, and education. We differentiate among these contexts according to distinctive sets of contextual informational norms governing data flows, to recognize the extent to which the integrity of users' interactions with news, health, and education websites is breached by advertisers. We trace user ID cookies that are used by trackers in more than one social context to understand whether trackers are potentially linking information about users against expected informational norms.

Previous studies on tracking across websites paid most attention to differences in the amount of tracking within rather than across social contexts. Researchers linked the amount of tracking within contexts to the incentives of publishers, but we try to understand the incentives of trackers to use persistent identifiers across social contexts. Our findings show that social contexts matter for trackers. Third of the observed third-parties had persistently identified users among social contexts. The usage of persistent user identifiers is taking place among all three social contexts, regardless of the sequence of browsing, and between each and every pair of websites under investigation. Trackers from the healthcare context are the most likely ones to link user identifiers across contexts. Specifically, users who visited popular news websites after browsing healthcare websites are more likely to be followed by trackers than users who read their news after browsing in educational websites. This contrasts previous findings that labeled healthcare websites as less risky for users' privacy. While tracking within healthcare websites might be marginal, the tendency of healthcare trackers to persistently identify users across social contexts is a serious threat to users' privacy. We also found that the usage of persistent identifiers across contexts is happening in different levels, and changes according to browsing sequence, with few websites that were found more prone than others to link user identities.

This is a first modest step to empirically apply contextual understanding to online tracking. We demonstrate how shedding light on cross-contextual tracking patterns reveals alarming privacy violations and calls for more empirical work on online tracking from a contextual perspective. Applying this framework to large-scale studies might open a new avenue for understanding unaccounted privacy disturbing practices by advertisers. Our findings show that current online privacy approach of notice & consent is definitely not shielding users against these kinds of industry practices.

Unaccounted Privacy Violation:

A Comparative Analysis of Persistent Identification of Users Across Social Contexts

By Ido Sivan-Sevilla, Wenyi Chu, Xiaoyu Liang, Helen Nissenbaum[1]

Ubiquitous tracking has become an inevitable part of our online environment. Behind our favorite websites, companies we may never have heard of are collecting data points on every aspect of our lives – our interests, purchases, health condition, locations, and more. These data points are then combined into exceptionally revealing behavioral profiles, exposing intimate parts of our identity and fuel the multi-billion-dollar advertising industry that claims to predict what we are likely to consume in order to target us with ads (IAB, 2019).

Specifically, inference on individual behavior across social contexts is critical for the profiling and targeting capacities of online advertisers. The industry takes advantage of the variety of social domains across websites to link information about individuals from contexts that hold different privacy expectations by users. For instance, when advertisers cross information about users' medical problems, educational interests, and news consumption habits they are in a position to better know when a user can be turned into a consumer and make purchasing decisions that advertisers would not be able to predict otherwise. Studies showed how data from different websites is aggregated and used to infer about the demographics and interests of users, exposing them to manipulative practices that try to make them click on the 'right' (personalized) advertisement at the 'right' (personalized) time (Barford et al., 2014; Bashir et al, 2019; Lecuyer et al., 2015). The advertising industry had defined these moments as 'prime vulnerability moments of consumers' (Rosen, 2013), in which users are 'uniquely receptive' (Google, 2016).

According to the theory of contextual integrity, the identification of these vulnerable moments by the advertising industry, based on information gleaned across social domains, fails to respect contextual informational norms, and in so doing, violates users' privacy expectations. Past work on online tracking examined client interactions with thousands of websites in bulk, providing an essential window into the widespread uses of tracking practices. These studies, however, had not provided a nuanced comparative understanding of how such tracking interacts with the distinctive informational norms associated with respective websites. Scholars highlighted the prevalence of certain trackers and the centralized nature of the advertising ecosystem but have not advanced our understanding of trackers work across contexts, linking users' activities from different social domains.

[1] Helen Nissenbaum and Ido Sivan-Sevilla gratefully acknowledge research support from: NSA: H98230-18-D-006 and NSF: CNS-1704527, SES-1650589, and CNS-1801501.

Comparative findings on tracking across websites are sparse, mainly revealing invasive tracking by news websites in comparison to other website categories and explaining those based on the incentives of news' publishers to monetize their products. But what can be learned from tracing tracking from the point of view of the trackers, across rather than within social contexts? Specifically, how do the interests of trackers explain the usage of persistent user identifiers across social domains? Are certain social contexts more valuable for trackers than others? Can we spot variance in the way that trackers follow individuals across different social contexts, against expected informational norms?

To answer these questions, we investigate tracking across fifteen popular websites from three different social contexts – news, health, education - and offer a close empirical examination, based on client-side interactions with websites, of the extent that trackers use persistent identifiers for users across social contexts. These identifiers persist not only between websites but also across social domains. We assume, based on previous works (e.g. Bashir, 2019), that once a user ID is linked by advertisers across different contexts, browsing history can be used to infer sensitive attributes about a person – health condition, desire to quit a job, political orientation, and etc. Even though we cannot observe how advertisers use data to decide on the best targeting method per user, we know that there is a natural incentive in this ecosystem to aggregate data in order to learn large fraction of user's history and potential future behavior for targeting purposes (e.g. Acar et al., 2014; McGuigan, 2019). Surprisingly under-studied, using persistent user IDs across social contexts seems to have a significant impact on advertisers' profiling and targeting capabilities. We assume that trackers will try and link together their observations of user interactions across social contexts, against expected informational norms.

Though an instrumented Firefox browser (openWPM), we study tracking practices in the top-five popular websites of the three social contexts, going beyond the landing page of each website, to realize how advertisers utilize the conflation of contextual informational norms to better target users. We ran six experiments for all the combinations of browsing sequences across the three distinct social domains and studied stateful tracking practices via HTTP cookies. For each experiment, we are matching cookie ids among contexts to realize which trackers use the same user IDs for every context. We did not simply assume that the presence of trackers in two different social contexts means that they persistently identify the user across those contexts. Instead, we looked for a valid evidence, i.e. the usage of the same cookie id across contexts, to assume persistent identification of users in different social domains.

We found that social contexts matter for trackers. Usage of persistent user identifiers is taking place among all three social contexts, regardless of the sequence of browsing, and between each and every pair of websites under investigation. Trackers from the healthcare context are found to be the most likely ones to link user identifiers across contexts. Specifically, users who visited popular news websites after browsing healthcare websites are more likely to be followed by trackers than users who read their news after browsing in educational websites. This contrasts previous findings that labeled healthcare

websites as less risky for users' privacy (e.g. Englehardt and Narayanan, 2016; Cahn et al., 2016). While tracking within healthcare websites might be marginal, the tendency of healthcare trackers to persistently identify users across social contexts is a serious threat on users' privacy.

Additionally, we show how specific pair of contexts are linked differently by trackers across popular websites. The usage of persistent user identifiers between news and healthcare contexts takes place more massively, usually by dozens of trackers. Between healthcare and education contexts, however, there are only few trackers that link user information across contexts. Still, they are present in every pair of websites from these two contexts as well. We also found certain websites that are significantly more involved in cross-contexts user identification than others.

This paper highlights an often-neglected aspect in the online tracking literature – tracking practices across social domains. We measure the extent that trackers use persistent user identifiers across social contexts and aim to offer first empirical evidence on the practices of the advertising industry across online contexts through the lens of contextual integrity. We argue that what matters is not only the number of trackers in each social context, but also how those trackers choose to use persistent user IDs in different social contexts and potentially merge information about users among contexts. Trackers completely disregard contextual norms and users' privacy expectations for potentially 'better' targeting capacities. This is a rather unaccounted and under-studied privacy violation.

The experience of the data subject here is beyond Kafkaesque. We uncover privacy violations that are additional to the well-studied problems of the lack of transparency in online tracking activities or the fact that very few actors track users across the majority of websites. We provide a glimpse on how trackers are utilizing the conflation of contextual informational norms and potentially link users' activities from different social contexts. This can significantly fuel the manipulation power of this industry and allow advertisers to covertly and substantially influence users' decision-making.

In the next section we fit our contribution to related work on online tracking. Then we describe our methods – research design, data collection, and data analysis. In the third section we show the results, and the last section discusses the implications of the results, limitations of the work, future research avenues, and conclusions from this study.

1 – RELATED WORK

Identifying users across websites is central for the ability of advertisers to link collected data points within and across multiple web pages to a single individual. There are two distinct approaches to generate unique identifiers for users. One is 'stateful,' where the client's browser saves identities locally, as a long string, usually via HTTP cookies or JavaScript APIs, and later retrieve these unique IDs to identify the same users across websites. The second type of ID generation is 'stateless,' and is

based on information about the browser and/or network to create a unique 'snapshot' fingerprint of the user in a given moment based on browser's type, canvas/font, web traffic, audio settings, and battery levels. These identifiers are not saved locally by clients' browsers but are observed and probably saved by trackers (e.g. Karaj et al., 2019; Englehardt and Narayanan, 2016; Yang Yue, 2020; Cahn et al., 2016).

The more websites that choose to integrate a particular third-party tracker, the greater the tracker's capacity to collect information about users across social contexts and construct more comprehensive user profiles (e.g. Binns et al., 2018). This allows trackers to show personalized ads based for users based on their past activity, behavior, and inferred interests (Englehardt and Narayanan, 2016; Urban et al., 2020).

Still, the extent of identifying users across social contexts has not received sufficient attention by researchers. Scholars traditionally choose to study tracking in bulk, inspecting information flows between a user's browser and thousands of websites, with no sensitivity to the different social contexts of each inspected information flow. Millions, and sometimes even billions (e.g. Karaj et al., 2019) of third-party content requests were examined by researchers to highlight the centralized nature of the advertising ecosystem and the increasingly significant presence of 'top-trackers' (e.g. Google, Facebook, Twitter) across the majority of measured web traffic (Libert and Binns, 2019; Libert, 2015; Englehardt and Narayanan, 2016; Solomos et al., 2019; Urban et al., 2020; Yang and Yue, 2020; Cahn et al., 2016; Lerner et al., 2016).

These studies, however, do not distinguish information gathered by trackers according to the social context it came from. We can assume that 'top-trackers' link user identities across social contexts but lack a comparative understanding of how using persistent user identifiers differs among contexts. For instance, do information flows from the healthcare context are more/less commonly linked to users' behavior in other social contexts? To what extent the fact that users browse to look for educational resources serve advertisers when the same users search news articles? We aim to conduct a context-sensitive analysis of identifying users across social contexts to gain a better understanding of how and to what extent trackers link information about users from different social contexts.

Web privacy scholars who do consider different segmentations of websites in their results show variance in the amount of tracking within each website category (i.e. health, finance, news), without paying attention to how tracking information flows among social contexts. An overarching finding from these efforts is that news websites contain significantly more tracking mechanisms than other types of websites (e.g. Englehardt and Narayanan, 2016; Binns and Libert, 2019; Karaj et al., 2019; Urban et al., 2020; Yang and Yue, 2020; Lerner et al., 2016). The difference in the amount of tracking for each category is usually associated with the incentives for publishers to include third-party trackers in their websites, rather than with the incentives of trackers to gather information from certain social contexts

in the first place. Different business models and funding resources across websites are hypothesized as a possible explanation for the observed variance of tracking amount across websites. Sites that lack funding sources and provide articles for free for instance, are pressured to monetize their websites with significantly more advertising (Englehardt and Narayanan, 2016).

But what about the interests of the trackers? Do they differently value the gathered information according to the social domain it was collected from and strategically choose when to persistently identify the user across social contexts? How likely are they to link user information from one social domain for the purpose of making decisions about the user in another social domain?

We argue that an equally interesting, often unaccounted, measure of web privacy is the extent that trackers use persistent identifiers for users among social contexts. This approach goes beyond a single website as the unit of analysis, and groups websites based on their associated social domains, to show whether and how user identifiers persist not only between websites but also across social domains. This approach is based on the theory of privacy as contextual integrity. Nissenbaum (2010 & 2019) argues that society is constituted by diverse domains that are defined based on certain functions, purposes and values. For instance, we have certain privacy expectations from our interactions with healthcare websites, but probably different privacy expectations from our interactions with news websites. In the healthcare context, our aim might be to get personally-oriented medical advice, and therefore, we choose to reveal information that we deem sensitive and expect that it will not be shared or used to make any decision about us beyond the healthcare context. In our interaction with news websites, however, our purpose is to consume publicly available news, and we are not required to provide intimate information about ourselves. Therefore, we hold different privacy expectations from those information flows and expect information from the healthcare flow to be kept separated from information about our interactions with news sites.

Indeed, according to the theory of contextual integrity, the mere fact that a tracker is involved in our interactions with healthcare websites might already violate our privacy expectations. This is a well-documented problem. We already know the extent of the inclusion of trackers in our interactions with websites within a given social domain (e.g. Urban et al., 2020). Unaccounted, however, is the extent that commercial actors utilize this privacy violation and not only 'pollute' information flows in a given social domain but also link flows from distinct social domains by persistently identifying users for the sole purpose of better targeting them with ads. Depending on the amount and diversity of collected information, targeting can become more or less specific, driving advertisers to bid high or low for serving an ad to a specific user (e.g. Olejnik et al., 2014).

Specifically, we map the information flows we wish to assess based on the five parameters from the theory of contextual integrity. We inspect information flows in which: (1) the sender of the information is always user's browser / the third-party tracker; (2) the receiver of the information is the tracker / the

user's browser; (3) the information type is a user ID that appears in context B but was already associated with the user in context A; (4) the transmission principle is notice & consent; and (5) the data subject is the individual who interacts with the website.

We argue that this type of information flows does not meet the user's privacy expectations, even though users probably agreed to such flow through accepting terms they were not able to fully understand (for problems in the notice & consent approach to users' information in online advertising see Barocas and Nissenbaum, 2009). Such flow provides the opportunity for the tracker to conflate two (or more) social contexts for which users are expecting strictly separated information flows. To make this more concrete, we do not want our interactions with healthcare websites to be coupled with our interactions with educational / news websites by actors who cannot support our health condition and for purposes that are beyond medical advice. But this is exactly what the advertising industry is doing – the fact that trackers can couple users' 'presence' (i.e. browsing history, interactions with each site) in websites from different social domains might allow better targeting for commercial purposes, but is a clear violation of users' privacy expectations.

Adopting the contextual lens and investigating tracking in websites based on social domains raise several socially important questions. For instance, is the usage of persistent user identifiers more prevalent between healthcare and news contexts, or between education and healthcare contexts? What is the most susceptible social domain for cross-context inference when users go online? Is information from healthcare websites likely to follow users more, to other social contexts, than information from educational websites? These questions deserve a close empirical investigation. We argue that as concerned users, we should be able to compare websites' privacy levels also upon the extent that our interactions in these respected contexts will follow us to other contexts for the purpose of 'better' targeting us with ads.

Previous attempts to empirically highlight privacy violations by the advertising industry were not sensitive to the social domains of inspected websites, and consequently, did not include specific measurement for the conflation of social contexts by trackers. Scholars addressed the total amount of observed tracking across websites or the lack of transparency in the collection and analysis of users' data as spotted violations of users' privacy (Libert and Binns, 2019; Englehardt and Narayanan, 2016; Karaj et al., 2016; Cahn et al., 2016). Yang and Yue (2020) specifically stated that trackers easily collect information on individuals while making it very difficult for users to figure out who they are by intentionally hiding trackers' organization information form the public. Other works studied the more hard-to-detect practice of 'cookie syncing,' in which trackers share user identities with each other, allowing tracker-to-tracker merges of user profiles (e.g Englehardt and Narayanan, 2016; Fouad et al., 2020; Acar et al., 2014). This practice was found to be pervasive (Papadopoulos et al., 2019) and Google, Facebook, Criteo.com, and Adnxs.com were found as the most prolific cookie-syncing parties

in the industry (Englehardt and Narayanan, 2016; Fouad et al., 2020). These studies, however, did not account for the extent that cookie syncing takes place across social contexts of the web. How server-to-server sharing of user IDs allow trackers to gather user information from different social domains? This is a missed opportunity to highlight how trackers conflate social contexts for the purpose of better targeting individuals.

In contrast, we aim to highlight exactly how and by whom the conflation of social domains is potentially taking place. Based on the theory of contextual integrity, we analyze how trackers utilize their presence in websites that constitute distinct social contexts to persistently identify the user and potentially serve more personalized ads. This tracking behavior is far from accepted by website users and requires a closer attention by web privacy scholars. The majority of users are 'not comfortable' with tracking and profiling, and do not want their data to be used for purposes other than proving the service they requested (RSA, 2019; McDonalnd and Cranor, 2014). They are especially uncomfortable with actors navigating their data between different contexts to get a better understanding of what users might be interested in (Inmoment, 2018). To the best of our knowledge, this work is the first attempt to pay close attention to online tracking across social contexts and uncover such privacy violation by the advertising industry.

2 – METHODOLOGY

To assess the usage of persistent user identifiers across social contexts by trackers, we decided to focus on websites from three contexts – news, healthcare, and education. Each context is constructed based on different purposes and values: user interactions with news websites are based on the willingness of users to consume updated information or knowledge, and therefore include low privacy expectations. Healthcare websites, however, provide users with medical advice and serve as a space for users to discuss their health concerns, applying higher privacy expectations on each information flows. Educational websites signal the interest and aspirations of users and serve as a space for users to explore and engage, leading for high privacy expectations as well.

For each context we chose the top-5 popular websites as ranked by Alexa.com (for a full list of browsed websites see Appendix A). The decision to choose popular websites was based on a finding from Yang and Yue (2020), according to which web tracking tends to occur more intensively on the higher ranked websites. Each website was associated to its respective context based on the services it aims to provide and the types of user engagement it is likely to create.

To assess the usage of persistent user identifiers by trackers among the three contexts we traced stateful tracking via HTTP cookies in six different experiments. We decided to focus on HTTP cookies in this study since they were found as (still) the most dominant technique to identify online users across websites (Roesner et al., 2012; Fouad, 2020). Our goal was to cover all the possible browsing sequences

between the three contexts and compare the number of actors that use the same user identities among contexts in each experiment. Therefore, we conducted six experiments, cleaning the browser's cookie storage between the experiments, but not during an experiment. We wanted to see to what extent trackers that are embedded in different contexts choose to use the same cookie ID in each of the contexts.

In addition to the landing page of every website, we randomly chose four inner-site pages per site, before the experiments, and consistently visited the same inner pages in each website throughout the experiments. We overall measured tracking in 75 webpages per experiment, in six experiments, interacting 450 times with the investigated publishers. Our decision to include inner pages in our tracking analysis stemmed from previous findings that reveal an increasing amount of tracking in those pages, specifically through cookies, compared to websites' landing pages (Englehardt and Narayanan, 2016; Samarasinghe and Mannan, 2019; Urban et al, 2020).

We are aware that cookie values are often hashed or encrypted when used by the same tracker in different domains (e.g. Fouad, 2020). We also acknowledge that the persistent identification of users is happening in server-side as well (e.g. Acar et al., 2014), in ways that are more challenging for detection. Hence, we expect our results to be considered as a lower bound on the amount of persistence identification of users across social contexts in the investigated websites.

3.1. Data Collection

The collect tracking information (i.e HTTP cookies, JavaScript Operations, and HTTP headers) of each browser interaction we used an open source-based automated web crawler – OpenWPM - that simulates real users' activity and records website responses, metadata, cookies used, and scripts executed (Englehardt and Narayanan, 2016). We performed stateful crawl and enforced the use of only one browser instance. During the crawling we did not set the "Do Not Track" flag and configured our browser to accept all 3rd-party cookies. We also used the "bot detection mitigation" to scroll randomly up and down visited pages. We set sleep time between inner pages to five seconds, and timeout between websites to 100 seconds. All the six experiments ran on AWS EC2 Ubuntu server on May 1st, 2020, and outputs were recorded in a SQLite database.

3.2. Data Analysis

We first observed the total number of distinct third parties that were embedded in the examined websites and interacted with the instrumented browser during the experiments. Since not all identified third-parties are necessary trackers or advertisers, we were matching their observed names with the whotracks.me database, an up-to-date data set that traces actors based on a browser extension installed by millions of users (Karaj et al., 2019). This method to recognize third-parties as trackers was already used by Urban et al. (2020).

Then, we detected all associated cookies with these third-party. We joined each third-party cookie entry into a (name, value) tuple and regarded each unique combination of name and value as a unique cookie pair. We grouped our data based on (cookie name, cookie value) pairs, selecting only those who appear in multiple contexts. We consider a cookie to be involved in cross-context tracking if it is known by at least one tracker in two different contexts.

Our next analytical assignment was to recognize cookies that are likely to represent user identifiers for their respected trackers. We applied a known methodology to recognize ID cookies (Englehardt and Narayanan, 2016; Acar et al., 2014): Ensured that the length of the value is larger than 7 and less than 101 and that the value remained the same throughout each experiment. We also verified that the value of the cookie between each two experiments is lower than 66% according to the Ratcliff-Obershelp algorithm (Black, 2004). Following Fouad et al. (2020), we do not put any boundary on the cookie's lifetime since domains can continuously update cookies with a short lifetime and map them on server side for a longer term of tracking.

3 – RESULTS

As described, we ran six experiments that differ on the sequence of browsing social contexts according to the following:

1. News → Health → Education
2. Health → News → Education
3. Education → News → Health
4. Education → Health → News
5. News → Education → Health
6. Health → Education → News

For each experiment, stateful tracking was enabled and we traced the cookie storage of our browser to analyze how the same cookie IDs were used among contexts. For the sake of simplicity, we will refer to third-party trackers that use the same cookie ID in two or more contexts as 'persistent identifiers.' For a list of all trackers that we recognized as 'persistent identifiers' during our all the experiments see Appendix B.

We observed activities of a total number of 1,121 third-party trackers that installed 891,772 cookies in our instrumented browser during the six experiments. On average, third of the third-party trackers observed in each experiment were categorized as 'persistent identifiers' (See Table 1).

	# of third-party cookies	**# of third-party trackers**	**# of trackers that are persistent identifiers [and percentage from total]**
Exp #1	175,891	187	57 [30%]
Exp #2	126,987	216	69 [32%]
Exp #3	141,517	187	40 [21%]
Exp #4	119,098	153	64 [42%]
Exp #5	153,352	170	55 [32%]
Exp #6	174,927	208	69 [33%]

Table 1: Descriptive Statistics of Observed Trackers, Cookies, and 'Persistent Identifiers' in each Experiment

We then wanted to observe how the extent of using the same cookie ID is changing according to the place of the context in the user's browsing sequence. First of all, we wanted to see whether certain trackers persistently identify users across all social contexts. Interestingly, we saw that only when websites from the education contexts are either first / second in the order (experiments #3-#6), persistent identification across all three contexts takes place. Table 2 presents the actors that were persistently identifying users across all social contexts and the experiments in which they were observed. Unsurprisingly, the well-known powerful trackers in the industry persistently identify users crosses all social domains, and that websites from the education context are significant drivers in the decision of trackers to persistently identify in subsequent contexts.

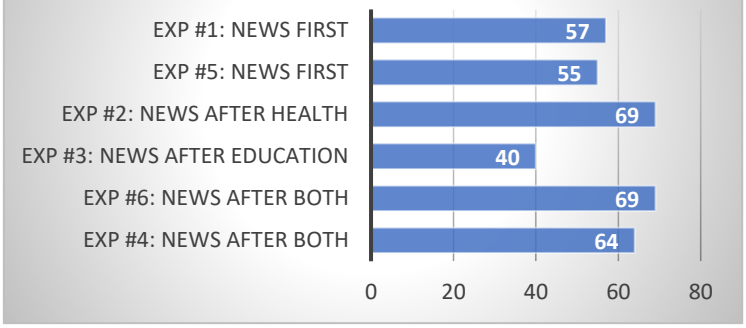
'Persistent Identifiers' Across all Three Contexts:	**Exp #1 H->N->E**	**Exp #2 N->H->E**	**Exp #3 E->N->H**	**Exp #4 E->H->N**	**Exp #5 N->E->H**	**Exp #6 H->E->N**
Doubleclick.net [Google]			V	V	V	V
Bing.com [Microsoft]			V	V	V	V
Twitter.com			V	V	V	V
Yahoo.com / Advertising.com [Verizon]			V	V		V
Facebook.com			V	V	V	V
Adsymptotic.com [Drawbridge Inc.]			V	V	V	V
Demdex.net/ Everesttech.net [Adobe]			V	V	V	V

Table 2: Persistent Identifiers Across all Social Contexts

Then we wanted to develop a more nuanced understanding of tracking across contexts by comparing number of persistent identifiers across experiments for each context.

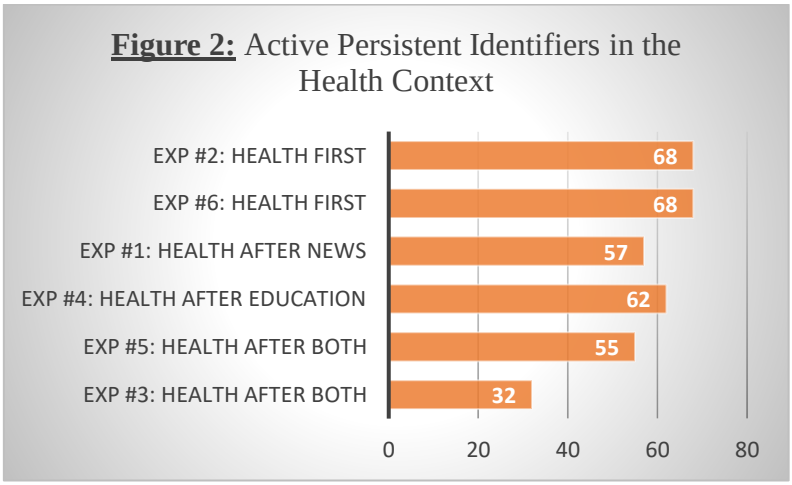
For the news context (Figure 1), the usage of persistent cookie IDs across contexts is happening by more trackers when prior browsing sites were from the health context rather than from the education context: 69 trackers have used the same cookie ID in the news context following browsing in the health context (experiment #2) vs. 40 trackers that did so following user browsing in educational websites. When browsing news websites takes place after both health and education contexts, the number of persistent identifiers is relatively the same – 64 and 69 actors.

Figure 1: Active Persistent Identifiers in the News Context



For the health context (Figure 2), this is not the case. The number of persistent identifiers stays the same, and relatively high, when browsing healthcare websites follows either news or education sites. However, websites from the healthcare context seem like a significant source for persistently identifying users. In experiments #2 and #6, when healthcare websites were visited first, we see the highest number of active persistent identifiers in this context (68 actors). When healthcare websites are the last one visited, this number drops to 32 (exp #3) and 55 (exp #5).

Figure 2: Active Persistent Identifiers in the Health Context



For the education context (Figure 3), we do not see a significant difference in the number of persistent identifiers based on the sequence of browsing. The numbers stay relatively low between experiments, and it seems that persistent identification takes place in this context by a rather lower number of trackers in comparison to the news and healthcare contexts.

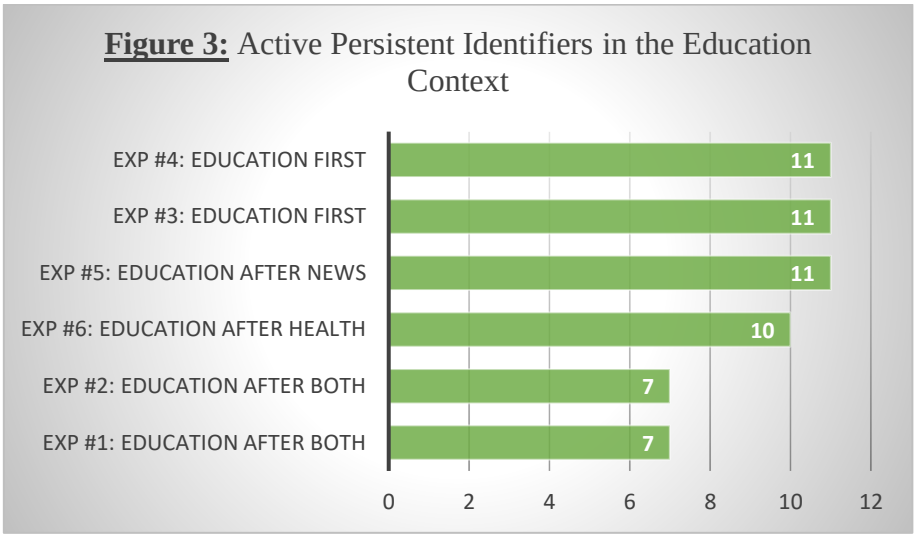


Figure 3: Active Persistent Identifiers in the Education Context

Our next analytical step was to zoom-in on each experiment and see how the patterns of persistent identification change per publisher (website). We traced each user ID cookie and created an edge between two websites in case an ID cookie was shared by a tracker that is present in both. We weighted each edge according to the number of trackers that share a user ID cookie between the two websites. We used Sankey diagrams to visualize the observed information flows (Figures 4-9). The width of each edge represents the number of distinct persistent identifiers that are active between contexts. We used Google Charts Library[2] to work with the plots.

Results show that certain websites are more prone to persistence identification than others and that the number of trackers that link user IDs between contexts varies significantly. Figure 4 below, visualizing our first experiment of browsing from news to health and then education contexts, shows that trackers in each and every examined news website are persistently identifying the user in all inspected healthcare websites, expect from TheGuardian.com, that shares a user ID only with Medscape.com. Some sites do so based on more trackers: washgintonpost.com -> webmd.com (17 actors); washingtonpost.com -> myfitnesspal.com (8 actors); frobes.com -> webmd.com (52 actors); forbes.com -> myfitnesspal.com (17 actors). Others news trackers share user IDs among healthcare websites via six trackers or less. Persistent identification from health to education follows a different pattern. Webmd.com and Mayoclinic.org trackers do not link the same user ID into education websites. Trackers from websites

[2] https://developers.google.com/chart/interactive/docs/gallery/sankey

who do persistently identify users - myfitnesspal.com, madscape.com, and verywellhealth.com - do so for all education websites via six or less trackers.

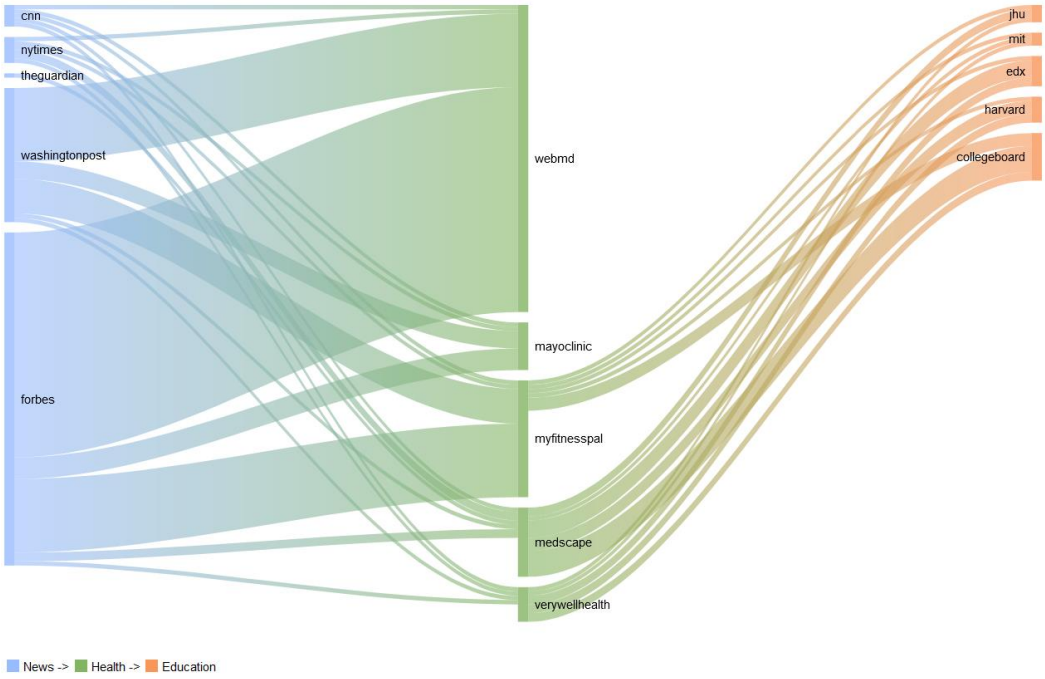


Figure 4: Persistent identification Flows in Experiment #1

Figure 5 below visualizes experiment #2, browsing from health to news and then to education contexts. In this experiment, we observe even more dense connections between healthcare and news websites, including for theguardian.com, which was less prominent in participating in persistent identification in the previous experiment. It seems that linking user identities from the healthcare to the news context is very appealing for news trackers. Persistent identification between news and education is happening at a much lower scale but is still taking place. Three news trackers from three different websites – cnn.com, nytimes.com and theguardian.com – persistently identify users in the education context, when users interact with collegeboard.org. Forbes.com, however, includes trackers that persistently identify users in all five education websites.

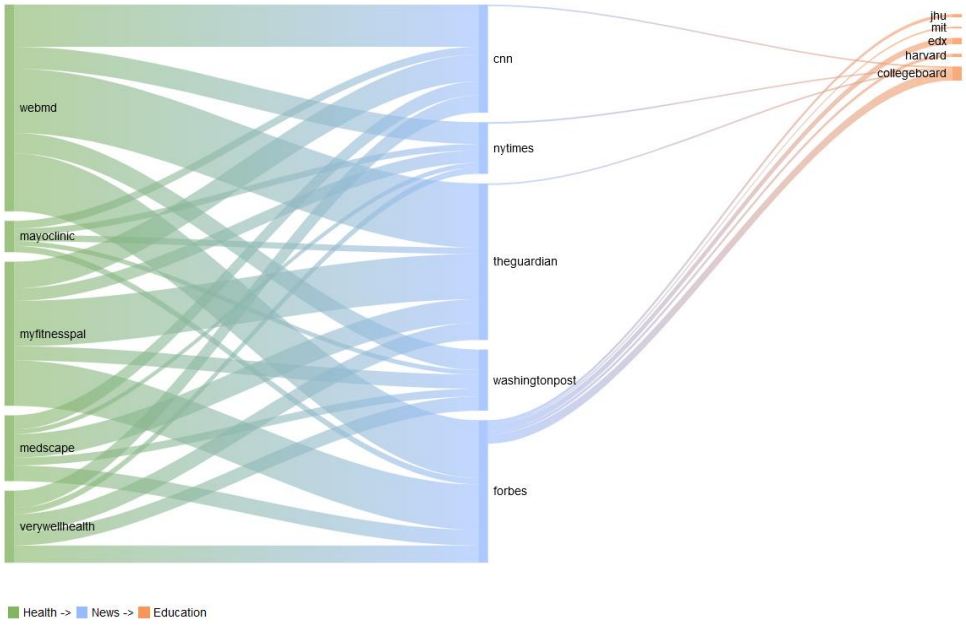


Figure 5: Persistent identification Flows in Experiment #2

Figure 6 below visualizes experiment #3, browsing from education to news and then health contexts. We see persistent identification flows mainly from edx.org and collegeboard.org. Noticeably, trackers embedded in mit.edu do not persistently identify users at all, into none of the other contexts. We also spot an additional persistent identification trend between contexts: trackers in all education websites, except from mit.edu, 'skip' linking to user IDs in the news context to persistently identify them only in the healthcare context, specifically when they interact with webmd.com. Persistent identification of users between news and health contexts is still relatively massive and takes place by the majority of persistent identifiers when users interact with webmd.com.

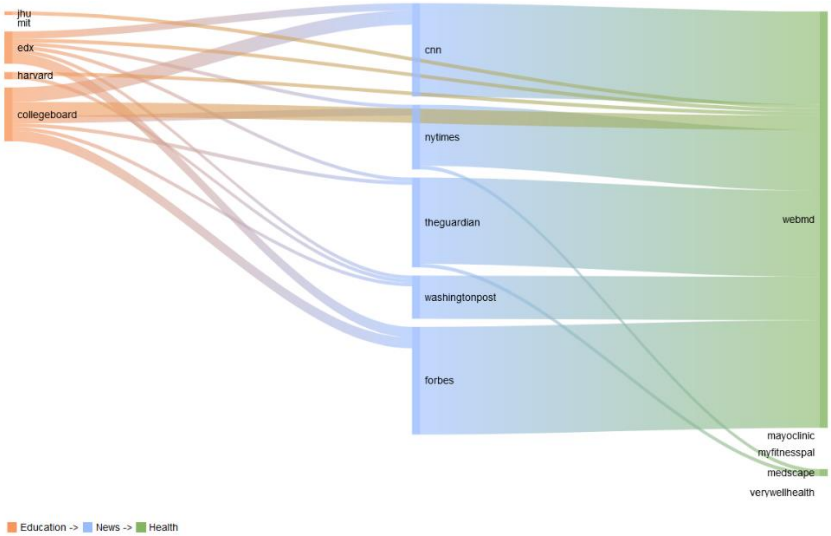


Figure 6: Persistent Identification Flows in Experiment #3

Figure 7 below visualizes experiment #4, browsing from education to health and then news contexts. The vast majority of persistent identification is happening from edx.org / Harvard.edu / collegeboard.com to other contexts. Again, mit.edu trackers do not participate in persistent identification, and there is one tracker that persistently identifies users from Johns Hopkins website (jhu.edu) to webmd.com. We see the trend of 'skipping' a context again, as edx.org and collegeboard.org trackers use the same user ID in the context of news consumption via forbes.com. Information flows of persistent identification from health to news are prominent, just like in experiment #2 (figure 5).

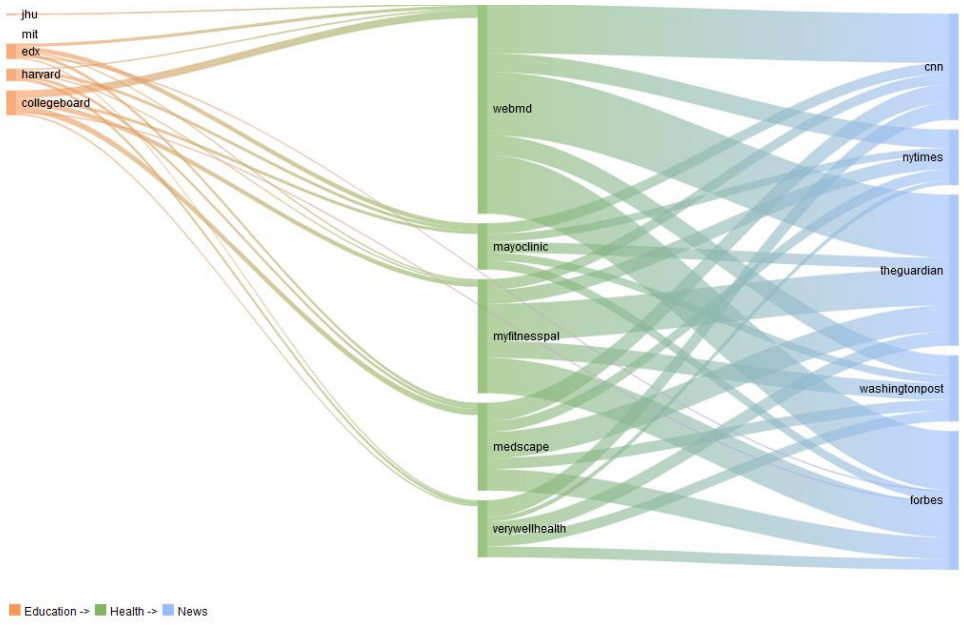


Figure 7: Persistent Iidentification Flows in Experiment #4

Figure 8 below visualizes experiment #5, browsing from news to education and then health contexts. It shows that trackers embedded in forbes.com are the only ones that persistently identify users from news to education contexts. Trackers in all other news websites 'skip' the education context and persistently identify users from the news to the healthcare contexts, via two healthcare websites – myfitnesspal.com and verywellhealth.com. Trackers in forbes.com are also exceptional in their number of persistent identifiers from news to health (e.g. 49 trackers in forbes.com persistently identify users in webmd.com). Trackers from the education contexts, across all websites, including mit.edu, persistently identify users when they interact with each and every healthcare website. Trackers from Collegeboard.org are doing so more often, but persistent identification is happening by trackers from all education websites, spotting user interactions in all examined healthcare websites.

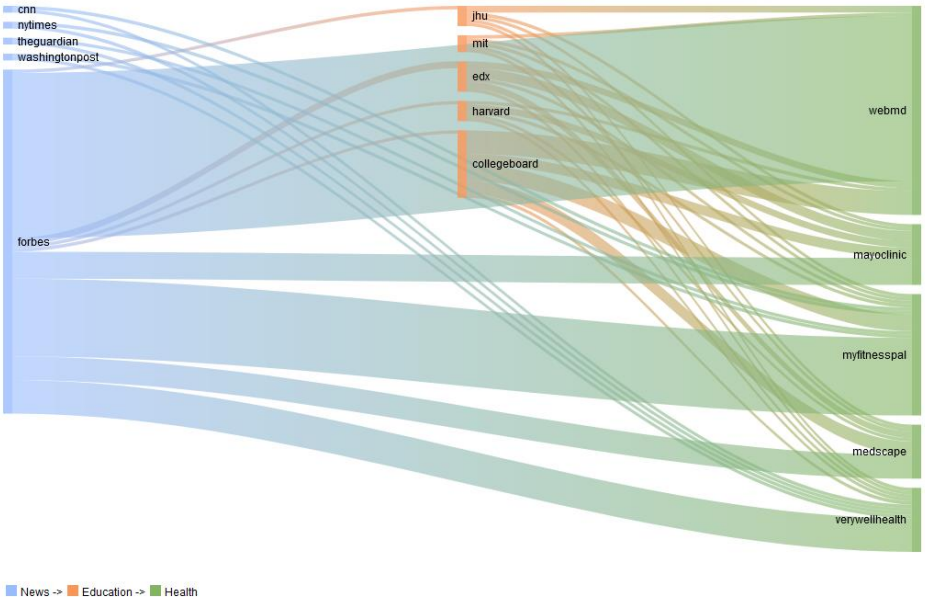


Figure 8: Persistent Identification Flows in Experiment #5

Figure 9 below visualizes experiment #6, browsing from health to education and then news contexts. Again, the strong persistent identification trend from health to education stands out. Trackers from each and every healthcare website persistently identify users in each and every examined education website. The same takes place from health and news contexts, albeit by a greater number of trackers from each news website. Interestingly, persistent identification trends from education to news are less significant than the ones observed between these contexts in experiment #3. Still, edx.org and collegeboard.org are the main source websites for such persistent identification.

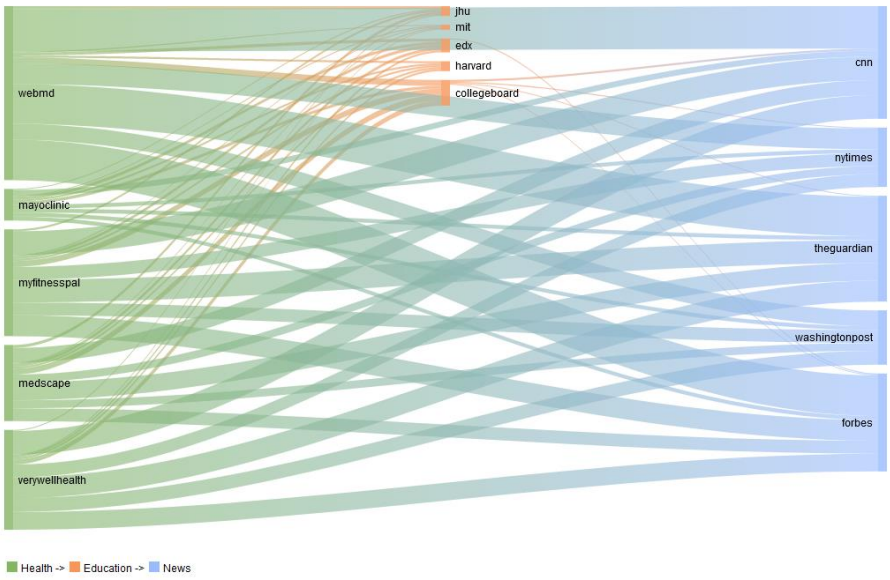


Figure 9: Persistent Identification Flows in Experiment #6

4 – DISCUSSION & CONCLUSION

Our results show that social contexts matter for trackers. Persistent identification of users is taking place between all studied social contexts, regardless of the sequence of browsing, and in each and every website that was under investigation. Thus, there is no single website, among the fifteen popular websites that were studied, for which trackers do not potentially utilize the conflation of contextual informational norms and try to link a user ID from on social context to another. On average, third of the observed third-party trackers were engaged in persistent identification in every experiment.

Specifically, for trackers associated with healthcare websites, and in continuation to findings from previous studies that showed low amount of tracking within the healthcare context (Englehardt and Narayanan, 2016; Cahn et al., 2016), we see that trackers associated with healthcare websites are the most significant drivers for persistent identification trends across contexts in our experiments. Simply put, users who consumed news or seek educational resources after browsing healthcare websites are more vulnerable to manipulation by the advertising industry. Comparing all studied contexts, healthcare-related trackers are more likely than trackers in other contexts to follow users in a stateful tracking mode. This makes sense from a tracker/advertiser point of view since health information is a valuable resource for manipulating and targeting users. This highlights that for users' privacy, what matters is not only the amount of tracking within a given social context, but also the extent of how trackers link information about users between the contexts for better targeting purposes. Healthcare websites, which were regarded by previous works as relatively less dangerous from a tracking perspective, are actually the most alarming ones, when it comes to persistent identification trends.

Trackers associated with specific websites in each context present different patterns of persistent identification. There are interesting variations in the patterns that link and persistently identify users across social contexts. Linking users between health and news contexts, some websites share many persistent identification instances by different trackers (e.g. forbes.com < --- > webmd.com for which 52 different trackers share user identities between the two sites), while others do so based on six or less trackers. It seems that linking user identities from the healthcare to the news context is very appealing for news trackers. Health and education contexts are also strongly linked, between almost every pair of websites from the two contexts, in both directions, albeit by a smaller number of trackers in every website.

Trackers from several websites stand-out in their alarming persistent identification practices among social contexts. Frobes.com was the only examined news websites that included trackers that linked news and into education contexts. These trackers were doing so for every studied educational website. Forbes.com also stand out in the scale of persistent identification, as it usually involves more than ten distinct trackers that are active between contexts in every experiment. Webmd.com stands out in the healthcare context. Trackers from education websites link user identities into the healthcare context

only through this site. Webmd.com also involves the most trackers that are engaged in persistent identification in the healthcare context. Finally, trackers from collegeboard.org are the most prominent persistent identifiers among examined educational websites, both in the variety of persistent identification links it creates, and the number of trackers that are involved.

Limitations

Limitations of this study are first of all the conservative approach that we took to spot persistent identification of users. As mentioned in section 2, these results should be perceived as a lower bound for the extent of persistent identification between the investigated contexts. Trackers often hash / encrypt / match user IDs in server sides in ways they we did not directly observe in our experiments. In addition, the physical location from which the crawl took place (New York, NY, US), as well as the dynamic nature of tracking practice that can change multiple times in a short period of time also limit the generality of our results. Also, we did not observe how user cookies are shared in URL parameters, in ways that allow third party trackers to share user IDs with another tracker. Even if different user IDs are assigned to different contexts, the linkage of IDs can be observed via HTTP requests in the browsing session. We did not include such inspection in our data analysis at the moment. Lastly, since we solely focus on stateful tracking, and as third-party cookies are increasingly blocked by browsers, persistent identification is most certainly taking place through means other than third-party HTTP cookies. These practices were not analyzed from our data yet.

Importantly though, we do not declare that this is an end of our contextual investigation journey of the advertising ecosystem. Vice Versa, this is just a beginning and a first modest step to empirically analyze the behavior of trackers across social contexts based on the theory of privacy as contextual integrity.

Future Work

Our goal is to continue and apply the theory of contextual integrity to empirically study tracking practices and possibly detect unobserved privacy violations in the wild. Looking ahead, we aim to understand how conflation of social contexts is taking place via JavaScript APIs, as the advertising industry is getting organized to the ‘post third-party cookie era’ (IAB Europe, 2020). We also seek to investigate cookie syncing of user IDs between third parties across social contexts to get a better sense of the amount of trackers who can potentially link information about users against expected privacy norms.

Conclusion

The web is an essential part of our lives, but an enormous source for tracking our activities and interactions across social contexts. This is a violation of our privacy expectations according to the theory

of contextual integrity. This study aims to empirically highlight such violations by considering the social contexts of users' websites for tracking analysis.

We have provided a first glimpse on how tracking is behaving differently in different social contexts. We showed how advertisers value healthcare data and increasingly deploy persistent identification practices from this context to others. Also, there are certain bonds between trackers who operate simultaneously in different social contexts (e.g. health and education) and utilize that to deploy practices against our privacy expectations.

Looking ahead, we aim to empirically uncover more of the conflation of contextual informational norms by the advertising industry, with a hope to remove the curtain between individuals and websites for better understanding of the 'invisible contracts' that we currently have with our digital service providers and ensuring they do not undermine our natural right to privacy.

In a 1999 New York Times article, Amitai Etzioni described cookies as 'surveillance files' and 'comprehensive privacy invaders' (Etzioni, 1999). Twenty-one years later, cookies (among other methods) are fueling a social norm of online life – the conflation of social contexts. Previous studies are highlighting the unbelievable quantity of tracking but pay less attention to tracking as it conflates used-to-be-separated social spaces. We should all work for keeping the integrity of our different social contexts when going online as well, not matter how profitable their conflation might be for certain parties.

Bibliography

Acar Gunes, Christian Eubank, Steven Englehardt, Marc Juarez, Arvind Narayanan, and Claudia Diaz. (2014). "The Web Never Forgets: Persistent Tracking Mechanisms in the Wild." *Proceedings of the ACM Conference on Computer and Communications Security (CCS)*.

Barford P., I. Canadi, D. Krushevskaja, Q. Ma, and S. Muthukrishnan. (2014) "Adscape: Harvesting and analyzing online display ads." *In Proceedings of World Wide Web*.

Barocas S. and H. Nissenbaum. (2009). "On Notice: The Trouble with Notice and Consent," Proceedings of the Engaging Data Forum: The First International Forum on the Application and Management of Personal Electronic Information .

Bashir M. A. (2019). *On the Privacy Implications of Real Time Bidding.* PhD Thesis, College of Computer and Information Science, Northeastern University.

Bashir M. A., U. Farooq, M. Shahid, M. F. Zaffar, and C. Wilson. (2019) "Quantity vs. Quality: Evaluating User Interest Profiles Using Ad Preference Managers*." In Proceedings of Networks and Distributed Systems Security.*

Binns R., J. Zhao, M. Van Kleek, N. Shadbolt. (2018). "Measuring third party tracker power across web and mobile". *ACM Transactions on Internet Technology (TOIT), Article 52*

Black Paul E. (2004). "Ratcliff/Obershelp pattern recognition", in *Dictionary of Algorithms and Data Structures* [online], Paul E. Black, ed. Available from: https://www.nist.gov/dads/HTML/ratcliffObershelp.html

Cahn A., S. Alfeld, P. Barford, and S. Muthukrishnan. (2016). "An empirical study of web cookies." In Proceedings of WWW, pp. 891-901.

Englehardt S. and A. Narayanan. (2016). "Online Tracking: A 1-million-site Measurement and Analysis." *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security,* pp. 1388-1401.

Etzioni E. (1999). "Privacy isn't dead yet." *The New York Times.* Available here.

Fouad I., N. Bielova, A. Legout, N. Sarafijanovic-Djukic. (2020). "Missed by Filter Lists: Detecting Unknown Third-Party Trackers with Invisible Pixels." Available in: https://arxiv.org/abs/1812.01514

Google. (2016). "The Basics of Micro-Moments." Available here.

Interactive Advertising Bureau (IAB). (2019). "Q1 2019 Reaches $28.4 B in US Digital Ad Revenues". *IAB.com.* Available here: https://www.iab.com/news/iab-advertising-revenue-q1-2019/

Interactive Advertising Bureau (IAB) Europe. (2020). "IAB Europe Launches Comprehensive Guide to Navigating the 'Post Third-Party Cookie Era'." *IAB.com.* Available here.

Inmoment. (2018). "What Brands Should Know About Creating Memorable Experiences." *2018 CX Trends Reports.* Available here.

Karaj, A., S. Macbeth, R. Berson, and J. M. Pujol. (2019). "WhoTracks .Me: Shedding light on the opaque world of online tracking." *arXiv:1804.08959v2.* Available here: https://arxiv.org/pdf/1804.08959.pdf

Lecuyer M., R. Spahn, Y. Spiliopolous, A. Chaintreau, R. Geambasu, and D. Hsu. (2015). "Sunlight: Fine-grained targeting detection at scale with statistical confidence." In *Proceedings of the 22nd ACM SIGSAC Conference on Computer and Communications.* Pp. 554-66.

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." *25th USENIX Security Symposium.*

Libert Timothy. (2015). "Exposing the Hidden Web: Third-Party HTTP Requests On One Million Websites." *International Journal of Communication* 9: 3544-61.

Libert T. and R. Binns. (2019). "Good News for People Who Love Bad News: Centralization, Privacy, and Transparency on US News Sites." *WebSci '19: Proceedings of the 10th ACM Conference on Web Science*, pp.155–64.

McDonald A. and L. F. Cranor. (2014). "Americans' attitudes about internet behavioral advertising practices." *WPES '10: Proceedings of the 9th annual ACM workshop on Privacy in the electronic society,* pp. 63-72.

McGuigan L. (2019). “Automating the audience commodity: The unacknowledged ancestry of programmatic advertising.” *New Media & Society* 21(11-12).

Nissenbaum H. (2010). *Privacy in Context: Technology, Policy, and the Integrity of Social Life*. Palo Alto, CA: Stanford University Press.

Nissenbaum H. (2019). “Contextual Integrity Up and Down the Data Food Chain.” *Theoretical Inquiries in Law* 20(1):221-56.

Olejnik L., T. Minh-Dung, and C. Castelluccia. (2014). “Selling off privacy at auction.” *In Proceedings of Networks and Distributed Systems Security.*

Papadopoulos Panagiotis, Nicolas Kourtellis, and Evangelos P. Markatos. (2019). “Cookie synchronization: Everything you always wanted to know but were afraid to ask.” *In The World Wide Web Conference, WWW 2019, San Francisco, CA, USA*, pages 1432–42.

Rosen R. J. (2013). “Is this the grossest advertising strategy of all times?” *The Atlantic.* Available here.

Roesner Franziska, Tadayoshi Kohno, and David Wetherall. (2012). “Detecting and defending against third-party tracking on the web.” In *Proceedings of the 9th USENIX Symposium on Networked Systems Design and Implementation, NSDI*, pages 155–168.

RSA. (2019). “RSA Data Privacy and Security Survey 2019.” *Rsa.com.* Available here.

Samarasinghe N. and M. Mannan. (2019). “Towards a global perspective on web tracking.” *Computers & Security 87*

Solomos K., P. Ilia, S. Ioannidis, N. Kourtellis. (2019). “Clash of the Trackers: Measuring the Evolution of the Online Tracking Ecosystem.” Available here.

Urban T., M. Degeling, T. Holz, N. Pohlmann. (2020). “Beyond the Front Page: Measuring Third Party Dynamics in the Field.” *WWW '20: Proceedings of The Web Conference 2020*, pp. 1275-86.

Yang Z. and C. Yue. (2020). “A Comparative Measurement Study of Web Tracking on Mobile and Desktop Environments.” *Proceedings on Privacy Enhancing Technologies* (2):24–44.

APPENDIX A – CRAWLED WEBSITES

News:

Cnn.com

https://www.cnn.com/politics

https://www.cnn.com/entertainment

https://www.cnn.com/style

https://www.cnn.com/health

Nytimes.com

https://www.nytimes.com/section/opinion

https://www.nytimes.com/section/health

https://www.nytimes.com/section/sports

https://www.nytimes.com/section/arts

Theguardian.com

https://www.theguardian.com/us/commentisfree

https://www.theguardian.com/us/culture

https://www.theguardian.com/us/lifeandstyle

https://www.theguardian.com/lifeandstyle/health-and-wellbeing

Washingtonpost.com

https://www.washingtonpost.com/politics/?nid=top_nav_politics

https://www.washingtonpost.com/opinions/?nid=top_nav_opinions

https://www.washingtonpost.com/business/technology/?nid=top_nav_tech

https://www.washingtonpost.com/world/?nid=top_nav_world

Forbes.com

https://www.forbes.com/money/#bc648bbc19aa

https://www.forbes.com/lifestyle/#3fb667f422d1

https://www.forbes.com/news/#12b68cc43690

https://www.forbes.com/leadership/#58e6b17b1d66

Health:

webmd.com

https://www.webmd.com/a-to-z-guides/common-topics

https://symptoms.webmd.com/default.htm

https://doctor.webmd.com/

https://www.webmd.com/family-pregnancy

mayoclinic.org

https://www.mayoclinic.org/patient-care-and-health-information

https://www.mayoclinic.org/departments-centers

https://www.mayoclinic.org/appointments/find-a-doctor

https://healthyliving.mayoclinic.org/?mc_id=global&utm_source=mayoclinicorgmainnav&utm_medium=l&utm_content=healthylivingprogram&utm_campaign=hlp&geo=national&placementsite=enterprise&cauid=100469

https://www.myfitnesspal.com/

https://www.myfitnesspal.com/exercise/lookup

https://www.myfitnesspal.com/food/search

https://blog.myfitnesspal.com/

https://community.myfitnesspal.com/en/categories

https://www.medscape.com/

https://www.medscape.com/familymedicine

https://reference.medscape.com/

https://www.medscape.com/diabetes-endocrinology

https://www.medscape.com/cardiology

https://www.verywellhealth.com/

https://www.verywellhealth.com/thyroid-test-analyzer-4178703

https://www.verywellhealth.com/lipid-test-analyzer-4582922

https://www.verywellhealth.com/cbc-test-analyzer-4768236

https://www.verywellhealth.com/renal-test-analyzer-4769439

Education:

https://www.jhu.edu/

https://www.jhu.edu/academics/

https://www.jhu.edu/admissions/graduate-admissions/

https://www.jhu.edu/research/

https://www.jhu.edu/life/

http://web.mit.edu/

http://web.mit.edu/research/

http://web.mit.edu/innovation/

http://web.mit.edu/campus-life/

http://web.mit.edu/admissions-aid/

https://www.edx.org/

https://www.edx.org/course

https://www.edx.org/schools-partners

https://programs.edx.org/professional-education/

https://www.edx.org/subjects

https://www.harvard.edu/

https://www.harvard.edu/faculty

https://www.harvard.edu/students

https://www.harvard.edu/admissions-aid#grad

https://www.harvard.edu/admissions-aid#titletop

https://www.collegeboard.org/

https://bigfuture.collegeboard.org/college-search

https://bigfuture.collegeboard.org/explore-careers/college-majors

https://bigfuture.collegeboard.org/pay-for-college/college-costs

https://bigfuture.collegeboard.org/get-in/applying

APPENDIX B – TRACKERS RECOGNIZED AS ‘PERSISTENT IDENTIFIERS’

News –

media.net,srv,crwdcntrl,rkdms,extend.tv,facebook,ipredictive,3lift,adform,bttrack,ib-ibi,rlcdn,exelator,bluekai,adsrvr,deepintent,casalemedia,amazon-adsystem,mathtag,sharethrough,openx,scorecardresearch,agkn,everesttech,bing,turn.com,bidr,tapad,w55c,sitescout,doubleclick,mookie1,adnxs,spotxchange,dnacdn,krxd,simpli,adentifi,taboola,1rx,twitter,zemanta,atdmt,adsymptotic,advertising,districtm,pubmatic,demdex,bidswitch,contextweb,mxptint,rubiconproject,myvisualiq,yahoo,owneriq,quantserve

Health –

media.net, srv, crwdcntrl, rkdms, extend.tv, facebook, ipredictive, 3lift, adform, bttrack, ib-ibi, rlcdn, exelator, bluekai, adsrvr, deepintent, casalemedia, amazon-adsystem, acuityplatform, mathtag, sharethrough, openx, scorecardresearch, appier, dotomi, agkn, everesttech, bing, turn.com, addthis, bidr, tapad, w55c, advangelists, sitescout, mfadsrvr, doubleclick, mookie1, spotxchange, adnxs, dnacdn, krxd, simpli, taboola, 1rx, twitter, zemanta, atdmt, tribalfusion, adsymptotic, advertising, districtm, rfihub, pubmatic, gumgum, demdex, bidswitch, zorosrv, contextweb, mxptint, rubiconproject, creativecdn, myvisualiq, iasds01, yahoo, owneriq, quantserve

Education –

doubleclick, demdex, everesttech, facebook, bing, linkedin, twitter, yahoo, atdmt, adsymptotic