FAST-TRACKING THE TRACKERS: APPREHENDING AND COMPREHENDING ADVERTISING TRACKING PRACTICES IN EDTECH

J. Bronfman, G. Kelly, J. Graham

Common Sense Media (UNITED STATES)

Abstract
The Common Sense Privacy Program has initiated a deeper look into advertising tracking practices in educational technology (Edtech). This paper will as an introduction begin by covering advertising tracking law and policy as it exists, including GDPR, Federal Trade Commission (FTC) jurisdiction, constitutional foundations, and federal legislation, as well as inroads states have made to protect consumers. The section of this paper covering types of advertising will further expose the difficulties of compliance with and enforcement of advertising law as it pertains to tracking. Following, this paper will explain the development methodology of the Privacy Program’s latest project to develop a tool that can evaluate over 3,000 Edtech-related company URLs to detect whether they use ad trackers. This section will describe the tools used, how we automated the detection process of third-party URLs on these 3,000 domains, and how to categorize which trackers would be considered advertising/tracking related uses based on deference to existing publicly available known advertising and tracker lists. Next, this paper describes the results of this research, including the technical challenges that may be encountered, planned analysis of the data gathered, and a summary of the industry’s gaps in transparency. Finally, this paper concludes with a hypothesis about the state of the ad tracker industry as it intersects with Edtech used by consumers and educators. This final section will address recommendations for each stakeholder group involved in the process, from developers interested in privacy by design, to regulators, educators, and consumers who interact with their designs.

Keywords: privacy, security, Edtech, students, apps, tracking, advertising.

1 INTRODUCTION

1.1 Why do we need to learn about advertising and trackers on webpages?
We need to learn about advertising and trackers on webpages primarily in the interest of transparency. Transparency, in and of itself, is a fundamental social value and legal principle underlying informed consent. Without informed consent, contract formation is suspect and, in some cases, void. This transparency, and its resultant informed consent, is the foundation of commercial, political, and personal speech online. The Common Sense Privacy Program is concerned that without this transparency, consumers, especially vulnerable consumers like kids, are not effectively consenting to the tracking occurring online. This concern has trained a spotlight on the need for this type of work regarding COPPA compliance and protecting children on the internet in general. We have identified a need for additional transparency for the practices of advertising tracking in the Edtech industry, and began this new research project to address this concern. As a result of this enhanced transparency, we also hope that Edtech and kid-focused products will improve their privacy, tracking, marketing and advertising practices.

1.2 What is Advertising/Marketing?
Advertising and marketing is at the onset the practice of selling goods and services. However, both as an integral part of this concept and independent of the selling of goods and services, advertising and marketing principles and practices are used in the “selling” of or promoting and suggesting ideas. With regard to children, this practice becomes more concerning. In this background section, a survey of existing research will illuminate the impact of trackers on privacy, particularly, for the Privacy Program’s purposes, the impact on student and children’s privacy.
1.3 Advertising Law

Advertising law is regulated in the United States by the Federal Trade Commission (FTC), a government agency tasked with commercial watchdog duties. The FTC is charged with regulating commercial behaviour and, in this instance, examining advertising to reduce advertising that is unfair or deceptive, and to promote advertising based on evidence. Commercial speech, i.e. advertising, does receive a lower level of protection than political speech, but it does still pass muster as protected speech under the first amendment of the U.S. Constitution that protects the rights of free speech, free assembly, and freedom of the press.[1] In the European Union (EU), advertising is regulated directly by the European Commission. Advertising subject to these regulations must not be misleading.[2] Further, advertising in the EU is indirectly governed by the General Data Protection Regulation (GDPR), including provisions covering data practices and transparency. GDPR recognizes the different roles associated with being a data controller versus a data processor, and the corresponding responsibilities with regard to data protection.[3]

1.4 Self-Regulation, Industry Standards, and Best Practices

In addition to official law and regulation, a variety of business practices imbed rules into the business of advertising and marketing in both offline and online worlds, with some permeability between the two worlds. Trade associations such as the 4As exist to support as well as monitor and police their own members.[4] For example, the 4A’s advertising association has been around for 100 years, but has standards and practices that currently apply to advertisers.[5] There is also a special category of self-regulation for children’s advertising.[6] Similar organizations exist in the EU, offering support and standards.[7] Companies that fail to meet these standards will have difficulty getting clients, respect and engagement from consumers, and possibly even draw the attention of regulators through complaints from these clients and consumers.

1.5 Types of Advertising and Relationship to Trackers:

While advertising (“ads”) as a common term has colloquial significance, for the purposes of this research we developed a more detailed taxonomy of the different types of advertising we intended to observe. The below explanations of data flow from party to party are a beginning of a structure to analyse advertising and tracking. Nevertheless, we still encountered some difficulties of definition and monitoring that must be highlighted. Contextual or otherwise called traditional advertisements display products and services to users based only on the relevant content or webpage in which the user is currently viewing, but contextual advertisements do not collect any specific information about the user in order to display these advertisements beyond a user’s search query or specific webpage they visited. Targeted advertisements, on the other hand, collect more generalized information about users from various sources that can include: demographic, location, gender, age, school, or interests. This information is collected in order to display products and services to a more specific target audience that may be more directed to users based on their information than simply contextual advertisements. The definition of targeted advertisements has been the subject of much discussion, but many state-specific student data privacy laws and federal laws protecting children prohibit targeted advertising based on any information collected from a user. Behavioural advertisements, on the other hand, collect more specific information about a user than targeted advertisements typically through the use of cookies, beacons, tracking pixels, persistent identifiers, or other tracking technologies in the product that provide more specific information about a user’s behaviour or activities over time. This behavioural or usage information can be used by the first-party vendor or shared with third-parties to display advertisements. These advertisements are even more persuasive to users than general targeted advertisements because the content of the advertisements is based on the highly specific information the first-party or third-party received from the user’s behaviour while using the application or service.

1.6 Tracking and Profiling

At the onset, not all trackers are exactly the same and we will need to make meaningful distinctions going forward. Trackers used for a number of different purposes including advertising, are found in several different formats, and appear in expected and unexpected places. For example, trackers can be present on different areas of an application or service (i.e., landing page, registration pages, login pages, or gated content pages) raising concerns about consumers expectations and often implying that the whole product is a safe place. On a more granular level, companies may think trackers are
fine on a landing or marketing page, but not login pages where Personally-Identifiable Information (PII) is collected with actual knowledge that the company’s users are children or students. However, the presence of trackers and persistent cookies on a website domain creates an environment where consumers are tracked across all pages, and they do not have adequate notice of where they should expect or not expect to find trackers. There are conundrums associated with previous research and tools intended to root out trackers in general, in the Edtech space, and to monitor their activities for the purposes of transparency, notice, and consent. There are also barriers to complete and effective transparency that should be explored, including structural, technical, and human factors. These barriers include, at the onset, (a) the burden of manual scanning and evaluation of URLs, (b) the different layers of ad tracking, flowing between multiple devices and vendors, and, of course, (c) the sheer quantity of websites overall, even when we narrow our focus to only Edtech products and services.

1.6.1 Understanding Tracking

In order to understand tracking, we have broken down tracking into its component parts and identified the parties involved in the process. Advertising tracking is the collection of information from users using persistent identifiers or other third-party scripts to recognize and track users on the first-party platform and across other third-party websites. Tracking of users allows the first or third party to display first or third-party advertisements to users on the first-party site or on other third-party sites across the Internet. There are relationships as well as disconnections. As an illustration of this process, we have identified the data flow relationships of advertising. In this simplified scenario, there are three companies, Company 1, Company 2, and Company 3. Company 1 can provide data to Company 2. Company 2 can provide data to Company 3. Company 1 may not know about Company 3. Company 3 may not know about Company 1. In this situation, we can begin to see how a user, and the initial company that has a relationship with that user, might not be able to control the flow of data.

1.6.2 Beyond Tracking

After completing the tracking process, the parties downstream from the tracking process may also engage in compiling and profiling. Compiling a profile of a user based on collection of non-personal information by the first-party for advertising purposes is the automated processing of personal data to evaluate certain personal aspects relating to a specific user. Profiling of a user can be used to analyse or predict aspects concerning that user for marketing or advertising purposes which is an indirect form of behavioural advertising. At this point in our research, we are focusing on the act of tracking, but future research may look at these downstream uses. We advise future researchers that, given current ability to access this information, these downstream uses may be found to be blackbox systems with limited to no regulatory oversight. Transparency so far has been minimal and would require that the companies voluntarily disclose their practices. Nevertheless, we encourage exploratory research in this area.

1.7 Marketing

In addition to advertising, parties involved in this process may also engage in marketing, which we have similarly broken down into categories and labeled the significant actors.

1.7.1 First-party marketing/First-party to USER (1ST -> USERS)

When a company that has a direct relationship with users on its own platform communicates new features, products, or services developed by the company. First-party direct marketing can be displayed as advertisements on the platform itself or communications sent to users by email, phone, text, physical mail, or other method. For example, a user has an account with a company that sends the user an email or physical mailer based on their personal information about a new product or service that the company has launched and the company thinks the user would be interested in learning more about it and purchasing the new product. The first-party receives a payment from the user if the user purchases their product.

1.7.2 Third-party marketing/Third to First-party (3rd -> 1st -> users)

When a first-party company communicates products from other third-party companies to users on its platform. The third-party receives no information about the first-party’s users. The first-party makes the decision which third-party products to communicate to which users. For example, a user has an account with a company that sends the user an email or physical mailer based on their personal
information about a new product or service a third-party company has launched and the company thinks the user would be interested in learning more about the third-party product and may purchase the new product. The first-party typically receives a commission or payment from the third-party if the user purchases the third-party’s product.

1.7.3 First to Third-party (1st -> 3rd -> users)

When a first-party company shares personal information of its users with third-party companies and allows those third-parties to send marketing communications directly to the first-party’s users. The third-party receives personal information from the first-party company about the first-party’s users. The third-party makes the decision which third-party or fourth-party products to send to which users. For example, a user has an account with a company that shares that user’s email address or other personal information with third-party companies that send marketing communications for third-party products to that email address. The first-party typically receives a commission or payment from the third-party if the user purchases their product, or a one-time payment for sharing user information.

1.8 Traditional advertising

1.8.1 First-party (1st -> users)

Similar to first-party marketing. When a first-party company displays first-party advertising to users on the company’s platform without using any personal information from users. For example, a user may or may not have an account with a company but that company does not use that user’s personal information to display advertising to the user about a new product or service the company has launched and the company thinks the user would be interested in learning more about it and purchasing the new product. The first-party receives a payment from the user if the user purchases their product.

1.8.2 Third to First-party (3rd -> 1st -> users)

Similar to third-to-first-party marketing. When a first-party company displays advertising from other third-party companies to users on its platform. The first-party does not use any personal information from users to display third-party traditional advertising. The third-party receives no information about the first-party’s users. The first-party makes the decision which third-party advertisements to display to which users. For example, a user may or may not have an account with a company, but that company does not use that user’s personal information to display advertising to the user about third-party products or services that the company thinks the user would be interested in learning more about it and purchasing the product. The first-party typically receives a commission or payment from the third-party if the user purchases their product.

1.8.3 First to Third-party (1st -> 3rd -> users)

Similar to first-to-third-party marketing. When a first-party displays to unknown individuals on third-party websites (such as CNN, Facebook, or Twitter), advertising based on demographic or non-personal information obtained from the third-party, to persuade that individual to engage with the first-party. The first-party has no personal information about the individual. The first-party makes the decision which third-party individuals should receive advertising of its products. For example, a company displays advertisements to potential users of its products on third-party websites about its first-party products or services that the company thinks the user would be interested in learning more about it and purchasing their product. The first-party receives a payment from the user if the user purchases their product. The third-party receives a payment from the first-party to display the first-party’s advertisements on their third-party service.

1.9 Behavioural Advertising

1.9.1 First-party (1st -> users)

When a first-party company displays first-party advertising to users on the company’s platform using collected personal information and specific behavioural usage information collected from the user’s use of the product to display advertising. For example, a user may or may not have an account with a company but that company uses a user’s behavioural usage information from the product to display advertising to the user about a new product or service the company has launched and the company
thinks the user would be interested in learning more about it based on their behaviour and will likely purchase the new product. The first-party receives a payment from the user if the user purchases their product.

1.9.2 Third to First-party (3rd -> 1st -> users)

When a first-party company displays advertising from other third-party companies to users on its platform. The first-party uses any personal or behavioural information collected from its users to display third-party targeted advertising. The third-party receives no information about the first-party's users. The first-party makes the decision which targeted third-party advertisements to display to which users. For example, a user may or may not have an account with a company but that company uses a user's behavioural usage information from the product to display advertising to the user about third-party products or services that the company thinks the user would be interested in learning more about it and purchasing the product. The first-party typically receives a commission or payment from the third-party if the user purchases their product.

1.9.3 First to Third-party (1st -> 3rd -> users)

When a third-party displays to first-party users on a third-party website (such as CNN, Facebook, or Twitter), behavioural advertising based on specific behavioural or usage information of the user obtained from the first-party to persuade that user to engage with the first-party again or with other fourth-parties. The first-party shares behavioural or personal information about the user with the third-party. The first-party makes the decision which first-party individuals on the third-party site should receive advertising of its products or other third-party advertisements. For example, a third-party displays advertisement to first-party users on third-party websites they visit about third-party products or services based on their behavioural usage information. The first-party typically receives a commission or payment from the third-party if the user purchases a product on their site. In addition, the first-party receives a payment for selling the user's behavioural information to the third-party. The third-party receives a payment from the user or commission from a fourth-party if the user purchases a product.

1.10 Ad Profile and Tracking

1.10.1 First Party (1st -> users)

When a first-party company creates persistent identifiers, compiles an ad profile, or uses other personal information from its users for tracking purposes on its platform to provide the features of its product.

1.10.2 First to Third-party (1st -> 3rd -> users)

When a first-party company shares persistent identifiers, ad profile, or other personal information of its users with a third-party company that serves first or third-party advertisements to those users based on the user's information on other websites across the internet. The first-party has personal information, or profile tracking information about the user that it shares with the third-party. The third-party may combine a user's tracking information with other tracking information it has acquired about that user from other third-parties and makes the decision which users should receive profile or targeted advertising.

1.10.3 Third to First-party (3rd -> 1st -> users)

When a third-party company shares persistent identifiers or other personal information of users to a first-party company to serve first or third-party advertisements to its users. The first-party has personal information about the user that it combines with tracking information about the user from the third-party. The first-party makes the decision which users should receive profile or targeted advertising.

2 METHODOLOGY

2.1 Research Description

In this research we used a list of over 3,200 URLs that we have previously used in our "State of Edtech Security Survey".[8] For each of these URLs we used an automated browser to navigate to the
respective URL we then captured a snapshot of all requests made after 10, 60, and 120 seconds. We used these intervals as they are approximations of the amount of time a visitor may spend on a particular site for either a low value site 10-20 seconds, an average page reading time for a session 60 seconds, and total time on a website 120 seconds. These are rough approximations that indicate industry averages. With this snapshot of requests we then used the ad-block engine used by the brave browser.[9] The ad-block engine allowed for the use of EasyList and EasyPrivacy, which are respectively used to identify requests related to advertising as well as requests made to trackers.[10] This process allowed us to get a sample of the amount of advertising and tracking requests that each respective web page initiates. In the event a site made multiple requests to the same URL that was either a known advertiser or tracker we counted each request separately. To facilitate making thousands of requests and the necessary processing GNU Parallel was used.[11]

2.2 Lists of Requests

In addition to the above data collection we also captured a list of all requests made across all sites to enable further analysis of advertisers and trackers across all sites. This allowed us to review how many different sites utilize a given tracker or advertiser. The below tables indicate the top 20 trackers and advertisers seen across all sites examined.

Table 1: Top 20 Advertisers Based on Number of Sites Making Advertising Requests to the Respective Domain.

<table>
<thead>
<tr>
<th>Advertiser</th>
<th>Number of Sites</th>
</tr>
</thead>
<tbody>
<tr>
<td>doubleclick</td>
<td>861</td>
</tr>
<tr>
<td>google</td>
<td>702</td>
</tr>
<tr>
<td>adnxs</td>
<td>271</td>
</tr>
<tr>
<td>linkedin</td>
<td>258</td>
</tr>
<tr>
<td>googleadservices</td>
<td>218</td>
</tr>
<tr>
<td>youtube</td>
<td>214</td>
</tr>
<tr>
<td>ads-twitter</td>
<td>200</td>
</tr>
<tr>
<td>rubiconproject</td>
<td>185</td>
</tr>
<tr>
<td>openx</td>
<td>171</td>
</tr>
<tr>
<td>yahoo</td>
<td>158</td>
</tr>
<tr>
<td>adsrvr</td>
<td>154</td>
</tr>
<tr>
<td>googlesyndication</td>
<td>151</td>
</tr>
<tr>
<td>pubmatic</td>
<td>147</td>
</tr>
<tr>
<td>casalemedia</td>
<td>144</td>
</tr>
<tr>
<td>bizographics</td>
<td>124</td>
</tr>
<tr>
<td>tapad</td>
<td>110</td>
</tr>
<tr>
<td>3lift</td>
<td>109</td>
</tr>
<tr>
<td>adroll</td>
<td>100</td>
</tr>
<tr>
<td>amazon-adsystem</td>
<td>89</td>
</tr>
<tr>
<td>spotxchange</td>
<td>70</td>
</tr>
</tbody>
</table>

Table 2: Top 20 Trackers Based on Number of Sites Making Advertising Requests to the Respective Domain.

<table>
<thead>
<tr>
<th>Tracker</th>
<th>Number of sites</th>
</tr>
</thead>
<tbody>
<tr>
<td>google-analytics</td>
<td>2354</td>
</tr>
<tr>
<td>doubleclick</td>
<td>1549</td>
</tr>
<tr>
<td>googletagmanager</td>
<td>1028</td>
</tr>
<tr>
<td>facebook</td>
<td>879</td>
</tr>
</tbody>
</table>
2.3 Examining Requests

As a preliminary assumption, we offered a baseline that industry norms are to make none or few advertising requests. This is a useful starting point as the ideal amount of advertiser requests in an EdTech application or context inclusive of children is zero. However, there are quite a few sites that make a non-normative and large amount of advertising requests. We observed that the majority of sites make either very few or no advertising requests.

2.4 Counting the Advertisers via Easylist

We used a resource called Easylist to create rules for tracking. Easylist is an industry standard that identifies advertisers and can be used to filter advertising from many web pages internationally, and also filter extraneous pictures, objects and other frames that surround the desired content. We decided to use this resource as it is the list that common advertising blockers use. There is significant visibility for these as we can view these in the web browser.

2.5 Counting the Trackers via EasyPrivacy

We also used Easy Privacy to identify tracking requests. By means of contrast, these trackers are more difficult to find and typically present no obvious indications that they are in use. These hide under a cloak of invisibility, hiding behind cookies, web beacons, and pixels. Only the most technologically savvy consumers will be able to view or control these trackers.

3 RESULTS

3.1 Summary of Key Findings

Our research indicated that tracking requests were much more prevalent on the sites we surveyed than advertising requests. With tracking requests, three-quarters of the sites used Google Analytics, and nearly half used DoubleClick. With regard to advertising requests about one-quarter of them used doubleclick. We found that 62% of the over 3,000 kid-intended sites we surveyed did not send advertising requests, but that means that 38% did send advertising requests. In addition, we found that only 12% of sites surveyed did not send tracking requests. As a result, 88% sent at least one tracking request. Further analysis yielded that among the 88% of the sites that sent tracking requests, 48% sent between one and nine tracking requests, and the other 40% of sites sent more than 10 tracking requests.
3.2 Process Observations

Due to the complexities of observational data there are several shortcomings with our process. Since this is an automated process, technologies to detect and flag so called bot-like or automated behaviour may result in our process not observing conditions that reflect what an average user may experience. Additionally, we only provide minimal page interaction and do not interact or otherwise engage with the site which in a normal setting may trigger additional requests to advertisers or trackers. For example, a user actually playing a game or learning using an educational module may behave such that more advertising or tracking requests are initiated.

3.3 Assumptions

This research relies on being able to identify advertising and tracking requests. With the increased usage of ad-blocking technologies there is an increasing incentive for the advertising and tracking industry to remain unobserved. As a result, there are several economic and public relations incentives for advertisers and trackers to avoid detection. As such our findings likely only identify a subset of the actual tracking and advertising that is happening.

3.4 Error in Identification

Additionally, there is the prospect of false negatives and false positives in our process for identifying trackers or advertisers. Managing these lists and ensuring they are narrowly focused is a substantial effort. Rather than assume authority in this realm we deferred to build on the tremendous work already done maintaining these lists. As such any questions of what constitutes an advertiser or a tracker are deferred to the methodology as defined by the EasyList and EasyPrivacy policy.[12]

3.5 Errors in Classification

In addition to the above limitations the detection library we are utilizing does not fully support all rules that may be present in EasyList or EasyPrivacy. The library creator indicates "Brave’s filtering library supports most of the rule formats mentioned that appear in EasyList, but not all of them."[13] As such there will be discrepancies between our methodology and error free classification of both advertising and tracking behaviour.

<table>
<thead>
<tr>
<th>Percent of sites making the respective advertising requests</th>
<th>Advertising Requests</th>
<th>Percent of Sites</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>62</td>
<td></td>
</tr>
<tr>
<td>1-3</td>
<td>11</td>
<td></td>
</tr>
<tr>
<td>4-10</td>
<td>15</td>
<td></td>
</tr>
<tr>
<td>10 or more</td>
<td>13</td>
<td></td>
</tr>
</tbody>
</table>

Table 3: Percent of Sites Making the Respective Advertising Requests
4 CONCLUSIONS

There are limitations, of course, with observational analysis. Specifically, there are limitations in detection via behavioural observation. Further, sites that detect that ad trackers may be observing automated bots on the other end and this may be potentially changing content delivery. Nevertheless, transparency in ad tracking is a worthwhile focus of research, in these beginning stages and beyond. We have taken a few more steps on the road to transparency and consumer education. Behavioural observation could take significantly more time and human input beyond the scope of the existing study, but perhaps not beyond additional prospective research involving human and/or artificial intelligence and machine learning involvement.

ACKNOWLEDGEMENTS

The authors wish to acknowledge the support and resources provided by Common Sense Media, inclusive of Common Sense Education and The Common Sense Privacy Program.[14]

REFERENCES


