Advertisement Network Analysis on Web Pages for Children

Annette Stawsky, Christopher Dare, Frank Rutaihwa, Rebecca Stevens, and Takuya Funahashi, *Carnegie Mellon University*

Abstract— **Online tracking is a prevalent practice that can be harmful when targeting children [5]. Online tracking can collect and share sensitive information such as name, date of birth, IP address, and sex. Many advertisers who use online tracking, develop profiles of inferred preferences and characteristics of users. They base these profiles on what the user searches for and what ads they interact with [13]. Advertisers use this information to create targeted ads, but the collection and use of sensitive information can pose privacy risks for all users. These risks are more acute for children because they are more influenciable and less experienced with being safe online [5]. This study does a large scale analysis of web tracking on children's sites. We filter our children's sites from 10,000 general sites, and compare tracking practices.**

We look at statistical differences between types of tracking (cookies, data requests), what percentage of the tracking attempts came from the website itself as opposed to third parties, whether these statistics vary across varying domains (educational vs entertainment mainly), which companies are doing the tracking and whether they are well-known. We intend to have COPPA in mind throughout the study and get a sense for whether children's websites are likely to be in compliance. If websites are in compliance with COPPA, we would expect to see statistically fewer third-party tracking attempts on children's websites. Our study finds the opposite, there is a larger average of cookies and requests on children's sites. Furthermore, a larger percentage of children's sites have tracking on them, and the third parties trackers don't vary greatly from general sites to children's sites.

Index Terms— **Privacy, Machine learning, Online Tracking, Children's Advertising, COPPA**

I. Background & Motivation

Web or online tracking is a process of recording, measuring, and analyzing individual behavior on the internet [8]. Due to the political, social and economical influence of the internet, user tracking has become ubiquitous [1][3]. Tracking can be used in a positive way to enable market segmentation and to adapt the message to the audience. But pervasively, it can prey on the audience to influence a behavior. Recent legal developments have woken up the possibility of regulation protecting online privacy [16]. However, tracking continues to be performed by first and third parties. Their collection, processing and analyzing of personal information without explicit consent can be seen as an invasion of privacy. Online, everyone is tracked and it is particularly dangerous on children's websites [5]. Children are especially susceptible to online dangers due to their lack of experience, and COPPA protects the personal information of children under 13 years old [16]. Web tracking allows access to browsing history which trackers can use to make inferences about a user's interests. Advertisers use these interests and machine learning techniques to create targeted ads. This study hopes to provide insight into the privacy ramifications of online tracking, especially as it relates to children. This insight can be guidance for regulators and policy makers on how to better protect user privacy online. Past studies identify general trends in web tracking, but have not done much in-depth analysis of websites that target children [8]. Therefore, we focus on tracking on children's websites and analyze how it compares to online tracking in general.

II. Related Works

Online tracking has been widely researched for years. Online tracking can be done with a variety of tools, and it affects an overwhelming percentage of websites. Of the top 10 sites returned by a search engine, more than 95% use tracking tools [4]. Roesner, Kohno, and Wetherall crawled the top 500 domains ranked by Alexa, and they revealed that more than 91% of those sites include more than one tracking tool [3]. HTTP cookies are the most well known tracking tool. According to a study by Englehardt et al, cookies can be used to trace browsing histories with a high accuracy of 62 to 73% [2]. Cookies were initially created to facilitate user interaction with a site. They allowed for functionality such as keeping track of which items were in a user's cart, or keeping a user logged in even if they closed the tab. The creator of the cookie, Lou Montulli, did not intend for it to be used for tracking and "confessed to mixed emotions about [that] development". In fact, he intentionally limited the information that a cookie could collect in order to avoid that possibility [13]. In our study, we will look at cookie trends on websites that target children as compared to general trends. We will also analyze another form of tracking known as data requests.

Tracking devices such as cookies are not only used by first-party websites, but also by third-party trackers. A study by Mayer and Mitchell shows how third-parties use tracking technologies such as HTTP cookies and device fingerprinting. They also develop a Mozilla extension called FourthParty that can be used for future web measurement [1]. When consent for tracking is legally required, companies simply craft lengthy privacy policies that no one bothers to read. On average it would take over 84 minutes to simply read the privacy policy for one site [12]. The intention behind legally required notice is to allow users to make informed choices about how their information is being used online. However, privacy notices are often too long and technical to be understood by someone who is not an expert in privacy and/or law. Even if such a person were to take the time to read the privacy policies, they often do not mention crucial information such as which third parties are receiving your sensitive information [12], [14].

In this way, web tracking is used in many online contexts and there are various discussions about its impact. The primary purpose of online tracking is for targeted advertising, so trackers will target groups of people that they can market to [13]. For example, the COVID-19 pandemic led to an increase in tracking on websites with COVID-19 information. A study by Tim Libert shows that 91% of general websites have tracking while 99% of websites with COVID-19 information have tracking. This study also found that even government and educational web pages had prevalent tracking, though slightly less than commercial web pages [15]. Just as there is sensitive information that should be protected such as health information, there are also vulnerable groups of people that should be protected. Recently, concerns have been raised about tracking users who have poor accountability and judgment: children. Zhao et al. report that children under the age of 11 are usually not fully aware of the risks of privacy [5]. In the U.S, there is a federal law to protect children from the potential harms of privacy violations. The law is called the Children's Online Privacy Protection Act (COPPA) which is a legal tool ensuring that caretakers can protect their childrens' privacy online [6]. COPPA protects the privacy of children under 13 years old from "commercial websites and online services." Among other requirements, this law requires website operators to do the following [16]:

- Inform children and parents of the privacy policy, and require consent from parents before collecting privacy information from children.
- Inform the parents whether they will share sensitive information with third parties, who the third parties are, and what sensitive information will be shared.
- Give parents the choice of allowing the website to collect sensitive information, and also prohibiting them from sharing that information with third parties.
- Allow parents to prevent the website from collecting or using personal information at any time.

However, operators (companies and organizations) are not always in compliance with COPPA. For example, HyperBeard, who is the developer of kids' mobile application, was sued for violating COPPA. HyperBeard tracked children without parental consent and was held accountable by the FTC [7]. Vlajic et al. analyzed 25 kids' sites, and found out that 50% of these pages contain tracking that is done through an invisible image [8]. There is also concern over automated bots/toys that engage in conversations with children, which might be collecting their data [11]. Tracking tools that are intentionally hidden raise questions of whether companies are really more careful/transparent with children's data.

A few prominent third-parties make up a large part of online tracking. A study by Englehardt and Narayanan on one million web pages highlights third party tracking trends. They rank the most prominent third-party and all of the top 5 third-parties belong to Google. Furthermore, 12 of the top 20 third-parties are also part of Google. In addition to Google, Facebook and AdNexus "are the only third parties present on more than 10% of sites" [9]. Part of our study will be to identify the prominent third parties on children's sites. According to a study by Gomer et al, third-party trackers vary depending on genre, geography, and the popularity of the site. For example, traffichaus.com and exoclick.com only appear in the adult category, and some trackers are connected with certain geographic locations [4]. This implies that there are different markets for different kinds of websites. One of the goals of this study is to analyze whether children's websites have a different market for third parties. Despite these findings, there is no large scale study of web tracking on children's sites. Vlajic et al. reports that tracking tools from Google-owned domain is the most frequently appeared in the sites for children [8]. Their study only manually analyzed 25 sites, which is not enough to see trends of web tracking in children's pages. In this study we will analyze a large dataset by classifying children's pages from a random 10k sample of 800k web pages.

To do analysis on children's sites, we considered a variety of previously implemented machine learning options. In order to filter out children's websites we considered measuring "word importance" with a few keywords that are correlated with children's sites. This can be done by applying term frequency-inverse document frequency (tf-idf) [10]. This gives a sparse data matrix corresponding to the words which are predictive of the content after removing common english stop words. Furthermore we looked into how to classify websites by their content. A study by Zhong Fan shows how supervised support vector machines and unsupervised K-means clustering can be used to study network traffic and classify it. They also find that these methods could give up to 95% accuracy [17].

III. Method

There are two components to our methodology: classification and analysis. The classification component consists of splitting our dataset of 10k websites into children's websites and non-children's websites. We further classify all domains depending on their top level domains as educational, government, europe, and others. The analysis component consists of comparing tracking statistics across groups, and within each group across domains. Our analysis was conducted using Python and we gathered statistics on cookies and data requests. The professors have been added as collaborators to our Github.

A. Scoring and Classification

First we filter all of the websites to extract those that are targeting children. We wanted to make this distinction because one of the goals of this study is to gain insight for COPPA compliance, and COPPA only applies to websites that are "directed to" children [16]. Furthermore, it is a common defense in FTC cases relating to COPPA that the website wasn't explicitly targeting children. This can be seen in the defense for HyperBeard, even though they produce games that are clearly intended for children [7].

We do this first level of classification with a simple filtering heuristic. We label every website that contains 'kid' or 'child' in the title as a children's website. We considered expanding this classification to include websites that contain 'kid' or 'child' in the description as well, but the results were too noisy to be useful. This filtering heuristic identified 77 children's sites.

We then tried to find clusters of the children websites. We do this using KMeans on the children's sites. Clustering into two groups, gave more accurate grouping than any other attempt. Viewing the clusters, we found that these children websites were either healthcare related or commerce related. We chose KMeans because previous studies have shown that unsupervised KMeans on network traffic can be up to 95% accurate [17].

We also categorized all the web pages into the following categories: educational, governmental, european, and others by analysing the top level domains. Upon seeing the unique top level domain names of our dataset, we realized that we could categorize the websites into these broad groups. We limited our categorization to the four chosen categories, because they were relatively easy to extract and categorize. Thus, analysing the tracking in these categories could help gain a better comparative understanding of web tracking.

Then we put the two categorizations together, thus a website could either have children focused in one of the two groups, or non children focused. Also, it can be from Europe, if it has European tld, governmental, educational or other. The other category consisted of all other top level domains that we could not classify as educational, governmental or European.

This classification brought to bare some few observations.

1. None of the websites categorized as children were educational. Or rather, the educational websites did not have words which our heuristic could flag them as children websites.
2. Government could have children websites. A website of treatment of children diseases by National Health Service, a UK health organ, would appear in our classification as children, in government category.

B. Analysis

Once the data is scored/classified, we are doing three levels of analysis. The first compares children's websites to non children's websites in general. This analysis will give possible insight into COPPA compliance and whether general trends change on children's websites. The second level of analysis is comparing within each group across domains. This will give insight about whether certain types of children's websites are more susceptible to privacy violations. The third level of analysis focuses on how third party trackers behave in all of the previous groups. Combined with the previous level, this might indicate what tracking on children's websites is being used for. If the prevalent third parties are not the same on children's websites, it could be that the purpose of the tracking is not for online advertising, or at least not on the same sites. We chose to analyze cookies, data requests, and prominent third parties because that is the type of analysis that other tracking studies have done [1] [12] [15].

We compute the following statistics:

- how many cookies/requests there are on any specific website. We use this to compare if the average amount of tracking decreases in children's websites as compared to websites in general. We also use this to see whether any specific domain is targeted more than others, and whether that same domain is targeted in the other group.
- how many third party cookies/requests there are on any specific website. We use this to see whether third parties are less interested in websites targeted at children (both whether the average number of third party cookies on a site is less, and whether the ratio of third party cookies to total cookies is less). We do a similar comparison across domains.
- which third parties are most prevalent. We capture this by comparing the cookie domain with the start domain of the website. If they are different, then the cookie originated from a third party. We use this to evaluate whether there is a different market for children's websites, and what the privacy implications of that might be.

Our data contains the start domain and the domain name for each cookie. We determine whether a cookie belongs to a third party if the domain name is not the same as the start domain. We determine the number of cookies on a given website by counting the entries for a given start domain, and we determine the frequency of a third party cookie by counting how often a cookie domain name appears in the entire set (across all start domains).

IV. Results

Using the classification heuristic mentioned above, we have classified 77 websites as children's websites. We found that the most prevalent top 20 cookies domains in general websites are similar to the top 20 of children websites. Figure 1 shows these top cookie domains on general sites, and Figure 2 shows the top cookie domains on children's sites. These results are somewhat surprising because we expected that there would be different third parties that target children. The only difference is the percentage of presence of a particular cookie domain, for instance doubleclick is available on more than 5.5 % of all websites, but only in 4.6% of all children websites.

The top 10 cookies domains in both categories are doubleclick, facebook, co

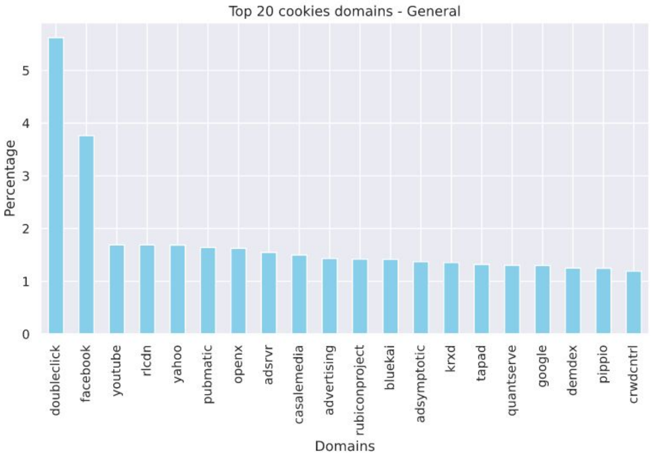

Figure 1: The 20 most prevalent third party cookies on normal websites.

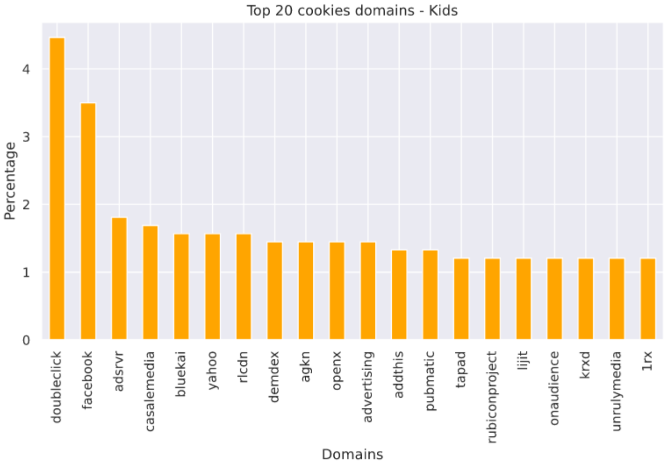

Figure 2: The 20 most prevalent third party cookies on children websites.

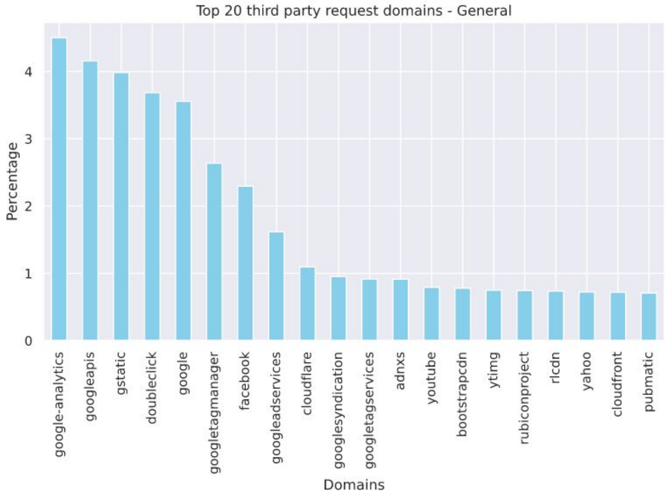

Figure 3: The 20 most prevalent third party request destinations on normal websites.

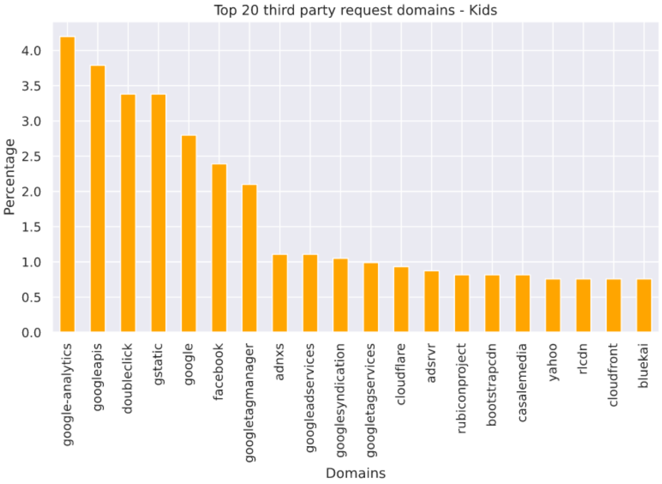

Figure 4: The 20 most prevalent third party request destinations on children websites.

We further found that there are slightly more third party cookies and requests to third parties in children's websites (Figure 5) as compared to all websites. We also saw that this was higher compared to european websites and educational websites.

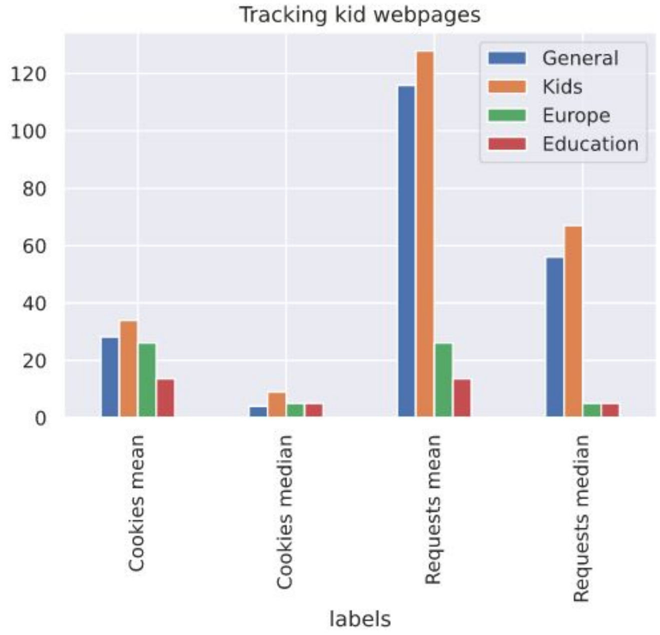

Figure 5: The average number of cookies and third party requests in children websites compared to all websites, european websites and educational websites.

It is likely that we are getting higher averages due to the fact that we have few children's websites in our sample. The appendix below shows some analysis of the websites when they are split into categories. They are described in the appendix because they are not the main findings, but they put the findings into context.

V. Discussion

Our results on the surface are somewhat surprising. The expectation, especially due to COPPA, would be that children's websites should have less third party cookies and requests than other sites. This is because in order for a website to be COPPA compliant, websites need to ask parent permission to not only collect sensitive data about a child, but also to share that data with third parties. Several cookies and requests do collect sensitive information about users, so it seems as though most of these cookies and requests on the children’s websites would need a permission associated with them. The higher number of cookies and requests on kids sites, especially from the third parties, looks as though it could be a COPPA violation on the surface.

Another surprise was the fact that markets for advertisers and data brokers on both types of websites was very similar. We expected the third parties used on children’s sites to differ from those on regular sites because of the difficulties associated with collecting data about children.

However, we cannot directly claim that the children’s websites we analyzed are violating COPPA. Firstly, most of the third party activity is coming from advertisers or trackers from Google. Most people have consented to allowing Google to track them at some point in their lives, whether intentionally or inadvertently, just due to their prevalence. We also have not looked at the privacy policies associated with the sites, and there is a chance that the third parties are included within it. When the privacy policy is agreed upon, often the permission for this data collection has been granted. Since many parents do not actually read the privacy policies, these trackers might have been overlooked. So despite our results showing the opposite of COPPAs intention, there is still a possibility that these sites are COPPA compliant.

Regulators would likely be surprised if we showed them this because it contradicts their intentions. With COPPA being a piece of legislation intended to limit the tracking of children online, the fact that there are more third party cookies associated with children's websites seems to suggest that either enforcement or requirements of COPPA should be altered. However, a regulator would look for more work to be done related to our research, as we faced many limitations.

VI. Limitations

There are a few limitations to this study and to the results. Given 10,000 websites, our filtering process only yielded 77 websites for children. Our filtering method could be improved to more accurately obtain children’s websites, such as filtering based on keywords related to kids (beyond just the words “children” or “kid”). This would give us a larger set of sites that are directed to children and we could manually filter them in case the keywords are too broad. Since we only had 77 children’s websites in consideration, our results are not likely statistically significant and this analysis should be applied to more children’s websites to see what happens on a wider scale.

There are also websites that are not directly targeted for children, but that children frequently visit. These are not included in our analysis, and are hard to detect based on keyword filtering. As younger kids become more and more competent on the internet, their ability to visit all sites increases, and it should be taken into account that many children are visiting websites that are not meant for their demographic. Youtube is a prominent example of this - even though there is a Youtube Kids, many children do not use this specific kids site.

Lastly, we only analyzed cookies and data requests, but there are many other forms of online tracking that could affect the accuracy of these results. However, this means that we are only underestimating the amount of tracking on children’s websites so our discussion still stands.

VII. Future Work

To build on our research, we would first improve the filtering heuristic and collect more children’s websites. That way, our results will have more statistical significance. We would also like to look into how exactly these websites go into obtaining parent permission and see how easy it is for a child to act as their own parent and accept the policy.

From a children’s privacy perspective, we can think of two useful continuations of this analysis. The reason we could not include them as part of our analysis is because they involved manually selecting a subset of our websites and interacting with them online. We did not have the time to interact with all of these websites individually throughout the semester. However, it would be interesting to 1. determine what kinds of information is being collected by trackers on children’s websites and 2. whether that information is mentioned in the privacy policy. In order to do the first part, we would have to select the most consequential websites to look into. It would make sense to include sites that have unusually high cookies/requests, and any that belong to a domain that is targeted at a higher rate. For example, Fandom wasn’t categorized as a children’s site for us but it has about 8000 cookies according to our measurements, so it would be an interesting place to start. Then we would interact with the site and look at the requests’ GET/POST payloads to see what information is being collected. If we see any sensitive information that we think should be mentioned explicitly, we scan the privacy policy to see if it is mentioned.

VIII. Conclusion

Online tracking can collect sensitive information that can damage user privacy. Our study shows that children are also affected, despite laws that are intended to protect their privacy online. We found that there are slightly more third party cookies and requests to third parties in children’s websites as compared to all websites. We also saw that this was higher compared to european websites and educational websites. We hope that this study paves the way for future research into how children can be protected online, and how regulators can enforce existing laws.

IX. Appendix

The following figures show some findings that we found interesting. They may help put our findings in perspective.

Figure 6 shows the average and median number of third party cookies and requests in children websites in commerce and healthcare domains. It is surprising that there are no big differences between the two categories, but slightly more in healthcare domains than commercial ones.

Figure 7 shows the top 20 third party request destinations by categories, and in comparison between all websites and those categorised as childrens'. Figure 8 shows a similar comparison using the cookies data.

Figure 9 and 10, show the websites with most third party cookies and request domains.

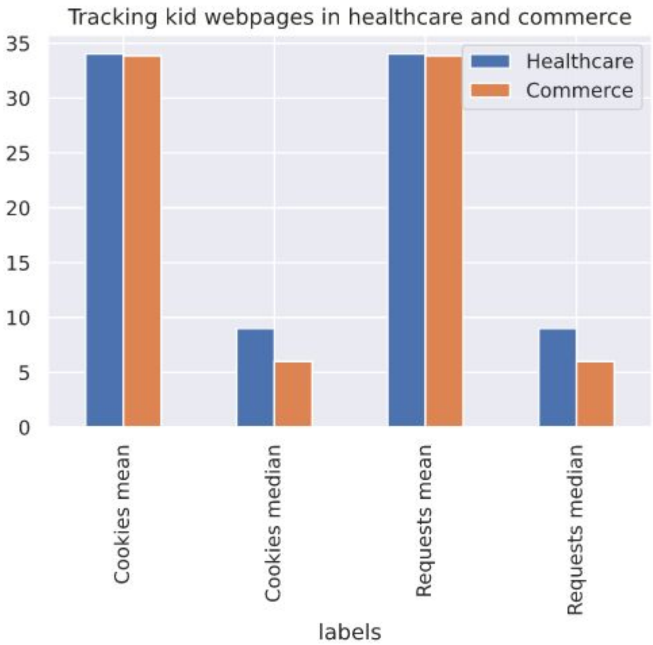

Figure 6: Number of third party cookies and third party requests in children websites belonging in healthcare and commercial domains.

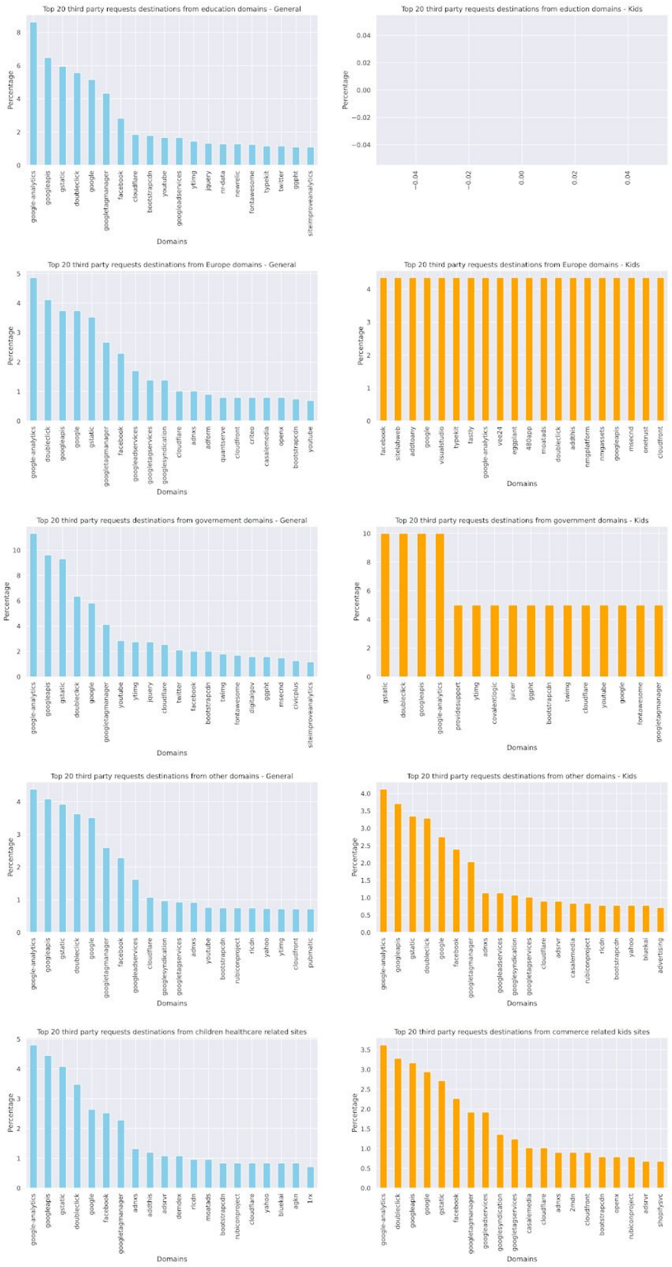

Figure 7: Top 20 third party request destinations, and their percentage of presence in the data of a particular category.

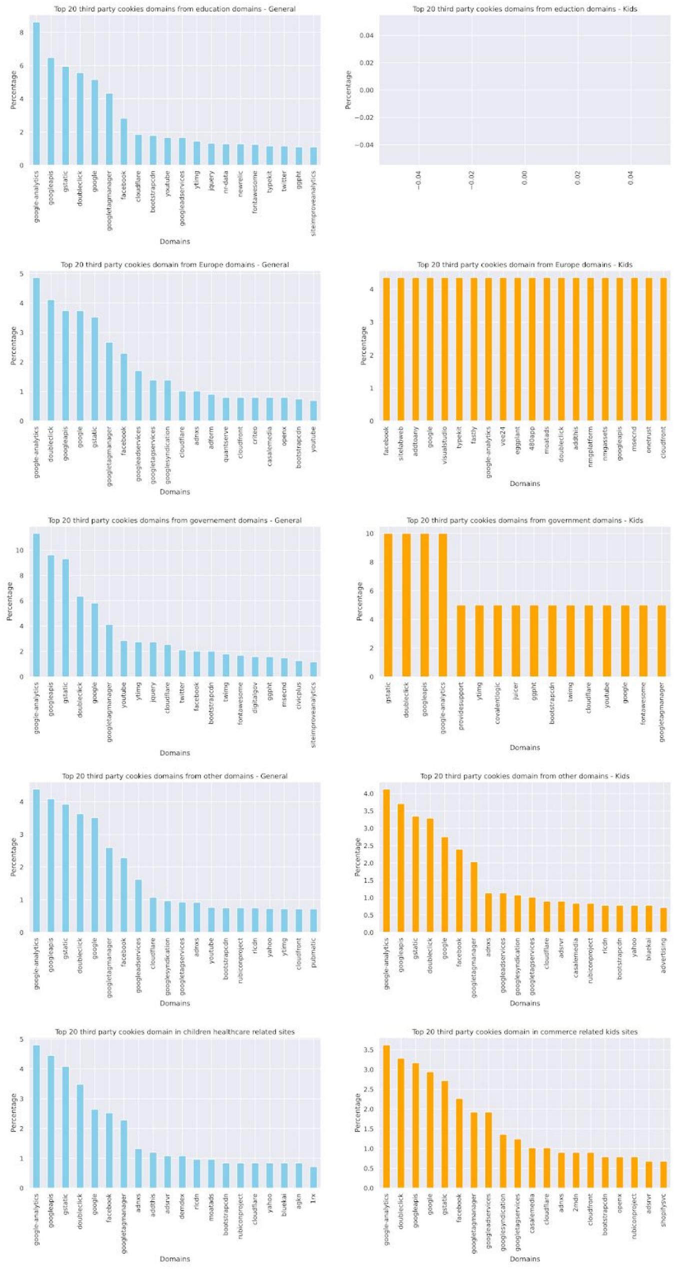

Figure 8: Top 20 third party cookie domains, and their percentage of presence in the data of a particular category.

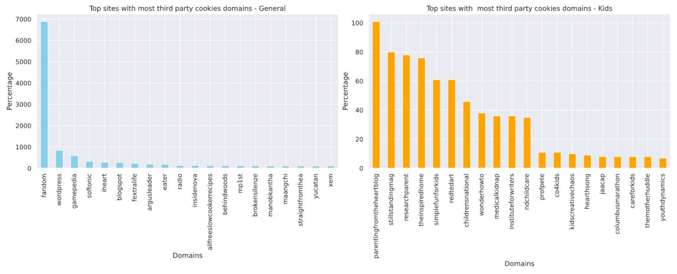

Figure 9: Websites with most third party cookies domain in general and for the children dataset.

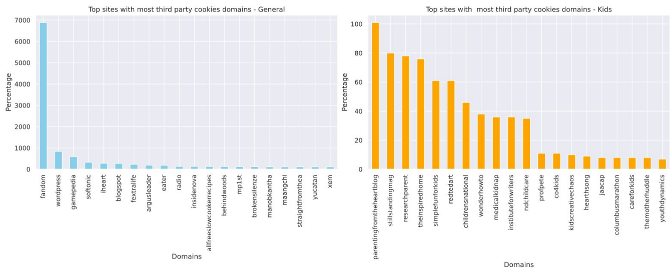

Figure 10: Websites with most third party request domains in general and in the children dataset.

References

[1] J. R. Mayer and J. C. Mitchell. “Third-party web tracking: Policy and technology”. In IEEE Symposium on Security and Privacy (S&P)), pages 413--427. IEEE, 2012.

[2] S. Englehardt, D. Reisman, C. Eubank, P. Zimmerman, J. Mayer, A. Narayanan, and E. W. Felten, “Cookies That Give You Away: The Surveillance Implications of Web Tracking,” in Proceedings of the 24th International Conference on World Wide Web (WWW '15). International World Wide Web Conferences Steering Committee, Republic and Canton of Geneva, CHE, pp.289–299, 2015.

[3] F. Roesner, T. Kohno, and D. Wetherall, "Detecting and defending against third-party tracking on the web," in NSDI'12: Proceedings of the 9th USENIX conference on Networked Systems Design and Implementation. Berkeley, CA, USA: USENIX Association, 2012, pp. 12-12.

[4] R. Gomer, E. M. Rodrigues, N. Milic-Frayling and M.C. Schraefel, “Network Analysis of Third Party Tracking: User Exposure to Tracking Cookies through Search," 2013 IEEE/WIC/ACM International Joint Conferences on Web Intelligence (WI) and Intelligent Agent Technologies (IAT), Atlanta, GA, 2013, pp. 549-556.

[5] J. Zhao, G. Wang, C. Dally, P. Slovak, J. Edbrooke-Childs, M, Van Kleek, and N. Shadbolt, "I make up a silly name’: Understanding children's perception of privacy risks online," in Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems -CHI '19, New York, NY, Paper 106, 1–13, 2019.

[6] "Children's Online Privacy Protection Rule ("COPPA")", Federal Trade Commission. [Online]. Available: https://www.ftc.gov/enforcement/rules/rulemaking-regulatory-reform-proceedings/childrens-online-privacy-protection-rule.

[7] United States of America v. HyperBeard, Inc., and Alexander Kozachenko and Antonio Uribe, Northern District of California, 3:20-cv-03683, Pending. [Online]. Available: https://www.ftc.gov/enforcement/cases-proceedings/192-3109/hyperbeard-inc

[8] N. Vlajic, M. El Masri, G. M. Riva, M. Barry, and D. Doran, “Online Tracking of Kids and Teens by Means of Invisible Images: COPPA vs. GDPR,” in Proceedings of the 2nd International Workshop on Multimedia Privacy and Security, New York, NY, 2018, pp. 96–103

[9] S. Englehardt and A. Narayanan, "Online Tracking: A 1-million-site Measurement and Analysis," in Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security -CCS’ 16, no. 1, pp. 1388-1401, 2016.

[10] A. Rajaraman and J. D. Ullman, *Mining of Massive Datasets*. Cambridge, UNITED KINGDOM: Cambridge University Press, 2011.

[11] G. Stevens, “Smart Toys and the Children’s Online Privacy Protection Act of 1998 Note,” *Smart Toys Child. Online Priv. Prot. Act 1998*, pp. 1–3, Jan. 2018.

[12] T. Libert. “An Automated Approach to Auditing Disclosure of ThirdParty Data Collection in Website Privacy Policies”. in Proceedings of the 2018 World Wide Web Conference. WWW ’18. Lyon, France: International World Wide Web Conferences Steering Committee, 2018, pp. 207– 216. isbn: 9781450356398. doi: 10.1145/3178876.3186087. url: https: //doi.org/10.1145/3178876.3186087.

[13] J. Turow. The daily you: *How the new advertising industry is defining your identity and your worth.* 2012. Yale University Press.

[14] Aleecia M McDonald and Lorrie Faith Cranor. 2008. “The Cost of reading privacy policies.” I/S: A Journal Of Law And Policy For The Information Society 4 (2008), 543.

[15] McCoy et al. “Prevalence of Third-Party Tracking on COVID-19–Related Web Pages” (2020).

[16] "Complying with COPPA: Frequently Asked Questions", Federal Trade Commission. [Online]. Available: https://www.ftc.gov/tips-advice/business-center/guidance/complying-coppa-frequently-asked-questions-0#A.%20General%20Questions

[17] Z. Fan and R. Liu, "Investigation of machine learning based network traffic classification," 2017 International Symposium on Wireless Communication Systems (ISWCS), Bologna, 2017, pp. 1-6, doi: 10.1109/ISWCS.2017.8108090.