The University of Iowa

Trackers & Browsing Speed: The Internet's Silent Rivalry

What is the Relationship between a Website's Trackers and its Respective Loading Time?

Lawrence Deng

Colin Hehn

Henry Krain

Angelo Zamba

Randy Zhang

CS:3640 Introduction to Networks and Their Applications

Professor Rishab Nithyanand

6 Dec. 2023

1. Introduction

Websites' tracking of user activities is a subject that is no stranger to the press and public eye, and it is an issue that is often associated with privacy concerns and commonly emphasized in VPN advertisements. For most end users, the impact of websites' data tracking appears miniscule, and as such the potential effects that this tracking may have on website loading time usually fly under the radar. Indeed, most businesses and site developers appear to perform a pretty solid job at optimizing said tracking, and for good reason: Models have estimated that businesses like Amazon lose approximately $1.6 billion dollars to just a second of increased webpage loading time (Eaton, 2012). Such models make it apparent that in the online ecosystem, every second, or even millisecond, counts. In this paper, we aim to understand the role which trackers play in influencing these precious seconds.

2. Our Research Question

We aim to answer the following research question:

- What is the relationship between a website's trackers and its respective loading time?

Our intention at the beginning of the study was to discover the correlation between the number of trackers a website has and its loading time, and the significance of said correlation, which gave us a pretty linear path of work: Detect how many trackers are on a site, find how long it takes to fully render, and repeat for several sites to achieve a large sample size. Unfortunately, it wasn't until after we had gotten the numbers that we realized some issues with this line of inquiry, costing us about half a week of work and research, and steering us towards the more general route of examining how the inclusion of trackers on a site at all affects its load time.

After getting tangible data and graphs generated for general website trackers and load time, we wanted to go one step further and include the effects of a more specific type of tracker on load time: the internet cookie. So while our research question will remain in the format seen here and on the title page, we will be including extra analysis in an effort to paint a more complete picture on the state of tracking and site rendering times. Yes, not all cookies are necessarily ‘tracking cookies,’ though the most common type of non-essential cookie on the web is the kind that tracks user activity for advertising or other related purposes. It is not irrational to believe that the cookies used to monitor users play some sort of stake in a webpage's loading time.

The Amazon sales research example mentioned sheds some light on why measuring the impact of trackers on loading speed would be important for the internet today, though it is approaching the subject from the limited lens of business. We wanted to understand the impacts of site loading times beyond just financial loss. When sites take longer to load, their "bounce rate" increases; in other words, people click off of the site before it renders more frequently than if the site had loaded faster. User tolerance of sites’ loading time tends to expire quickly past a certain point, due to their frustration with what is considered poor service. This means less sales in Amazon’s case, but it also reduces the website's PageRank, a metric used by Google's search engine to choose which sites the user sees first, resulting in less traffic for the site. This concept does have business implications, though hobbyists, bloggers, and many other types of internet publishers are directly affected as well; longer load times means less visitors, and few sites would call that desirable. Additionally, if users can better understand how trackers impact website load time, neither of which are particularly liked in abundance, they can make more informed decisions to improve their browsing experience online.

3. Methodology

To find the relationship between a website's trackers and its load times, we originally devised to have a single web crawler parse through the 1000 most popular websites, as listed in top_1000_urls.txt, collecting the number of trackers on each site along with the site's loading time. The crawler was implemented in cs3640_playwright_crawler.py with Playwright, and ran with a Chromium browser on one of our personal machines. We made the decision to timeout any site that did not respond after 30 seconds, since this seemed an appropriate threshold for preventing misleading data. The crawler was able to successfully collect data on 642 different URLs, and recorded its findings in DataTrackerResults.csv.

However, after examining the data from this first web crawl we found it to be unsatisfactory in answering our research question. We thus devised two more web crawls to better understand the impact of trackers on site loading time. The first new crawl that we devised was identical in design to our original web crawl, with the critical difference that we ran it on a Chromium browser with the Ghostery extension installed. This extension blocks tags and trackers embedded in web pages, which gives us an upper bound of how quickly a website loads without its trackers. We implemented this crawler in cs3640_playwright_crawler_notrackers.py, and again ran it on a personal machine. The crawl managed to collect data on 617 URLs, with 576 of these URLs also being present in the first crawl's dataset.

The second new crawl, and the final one we chose to perform, examined cookies specifically instead of trackers in general. This crawl was implemented with Selenium rather than Playwright, and collected data on the number of cookies and total size of cookies, in bytes, on each site in top_1000_urls.txt. The crawl was able to obtain data on 838 URLs and wrote

them to DataCookieResults.csv. We then merged the intersection of this crawl's dataset with the original crawl's data set, resulting in 563 data points in DataMergedCookiesAndTrackers.csv, which contains both the loading times of the websites (with trackers) and the amount and size of cookies on each site.

After performing the three crawls, we graphed the resulting data from the CSV files and ran regression tests to observe any correlations present and determine how strongly the presence of trackers, number of trackers, cookie size, and/or number of cookies affect the load time of a website. These tests would allow us to better analyze the trends that exist within our collected data, and the graphs would allow us to better visualize how our predictor variables (trackers, cookies, and size of cookies) relate to our response variable (load time).

While the data we obtained and analyses we performed were useful in examining the effect of our predictors on website load time, extraneous variables existed within the study that we were not able to control, such as network speed and server location, which could impact the resulting load times of the websites. To minimize their influence, we created a final graph that looks at each website that we parsed and compares the load time for when we used a normal Chrome browser to when we used a Chrome browser with the Ghostery extension to block trackers. This approach allows us to see how trackers play a role in websites load time while reducing the extraneous variables that can affect the result. We then took the average of the difference between load times to find the overall change. If the average difference was relatively small, the relationship between trackers and load time would be considered insignificant; however, a large magnitude of difference between the two crawler datasets would indicate that trackers affect the load time of a web page significantly.

There are a multitude of ways we could have answered our research question besides the method we chose to pursue. Rather than looking at real webpages, which are subject to a lot of uncontrolled variables, we could have created a set of custom websites with varying trackers, cookies, and cookie sizes and then running our crawler over these dummy websites and collecting our predictor variables and load times for these custom-built websites. This would provide a more controlled environment where we could isolate specific variables while reducing the amount of outside noise that could interfere with our results. We also could have collected data using alternative browsers like Brave, Firefox, Edge, or Safari to see if trackers and load time differed based on the browser, which would have made our study more comprehensive. Similarly, we could have used a different tracker-blocking extension besides Ghostery to observe differences in tracker-blocking. Furthermore, the study could have focused more on the types of trackers, creating a crawler that collects data for different trackers, using common tracker attributes, and seeing which trackers affect load time the most. This would allow us to run multiple linear regressions and see which types of trackers play a larger role relative to the others.

While we believe our methodology and approach worked well, it was limited for a variety of reasons. Our approach only examined the websites with each crawler one time. It would have been better to parse through the websites multiple times with the same crawler and calculate the average predictor variable and load time before graphing and running a linear regression to account for the potential variability in each crawl. This would make any outliers in our initial sample less impactful in our results. Another limitation was that we did not account for other variables like website server location or network performance. A website with an American server would most likely load faster than a website with a Russian server simply due

to the distance between the server and client. A better approach would be separating the websites by regions to minimize these extraneous variables. We also could have been less prone to error if we collected all available data in one crawler instead of using multiple crawlers for each predictor variable in our domain. This would be more efficient and collect the data in the same environment, which would reduce possible errors due to differences in their environment.

4. Results

In performing our web crawls and and collecting data on trackers and website loading times, we aimed to find evidence answering three specific aspects of our larger research question:

1. Is there evidence of a strong correlation between the number of trackers a site contains and its loading time? That is, is the number of trackers a site has a good predictor of the time it takes to load?
2. Is there a significant difference in the loading time of a website when it has its trackers blocked by a browser extension compared to when its trackers are not blocked?
3. How does the number and total size of cookies on a website impact its loading time? Does it have a stronger or weaker correlation to website load time compared to looking at the number of trackers on a site?

With the data gathered from our three web crawls discussed in Section 3, we present the results of our study and evaluate the impact of trackers on website loading time, identifying any significant features that we come across in our evaluation.

Number of Trackers and Loading Time. Our very first crawl, measuring the number of trackers against each website's loading time, provided evidence for only a very weak relationship

between the two variables with many anomalous data points in both the predictor variable space (number of trackers) and the response variable space (loading time), with a correlation coefficient of only 0.12 despite a statistically significant P-value of 0.000244. The plot of the dataset is displayed in Figure 1 below.

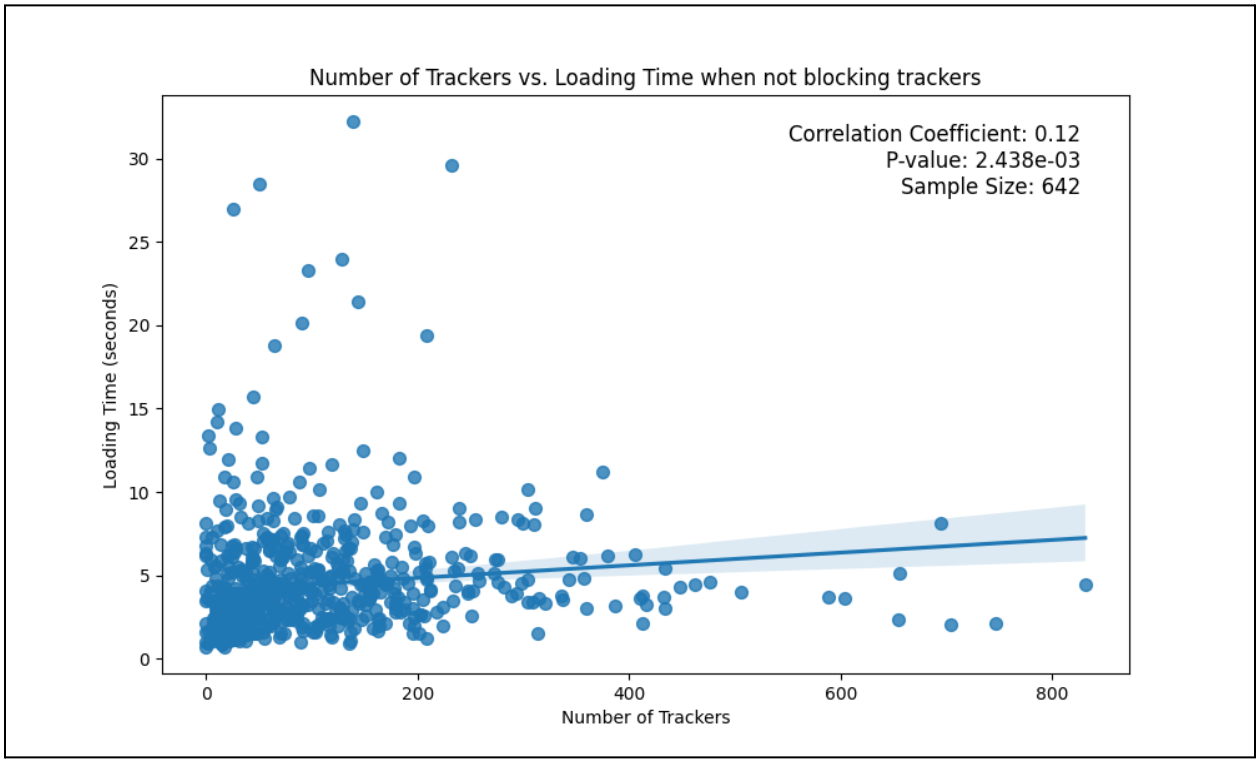

Figure 1. Graph of relationship between number of trackers versus loading time of websites

Even without the anomalous observations, the vast majority of data points were sites with less than 400 trackers and loading times less than 10 seconds, with no strong trend within those ranges. This suggests that the number of trackers is not a good predictor of loading time, and that its impact is small in comparison to other factors such as server location and network speed.

Load Time of Blocked vs. Unblocked Trackers. In our second crawl, we again looked at the number of trackers of a website versus its loading time, but in this crawl we used the Ghostery

extension with our Playwright crawler to block the trackers on each site we visited. Since the results of our first crawl did not indicate a strong relationship between number of trackers and loading time of a website when trackers were not blocked, we also do not expect there to be a strong relationship between the two variables in the resulting dataset. Our data supports this claim, as we once again find a weak correlation between the two variables in the websites that we crawled, as shown in Figure 2 below.

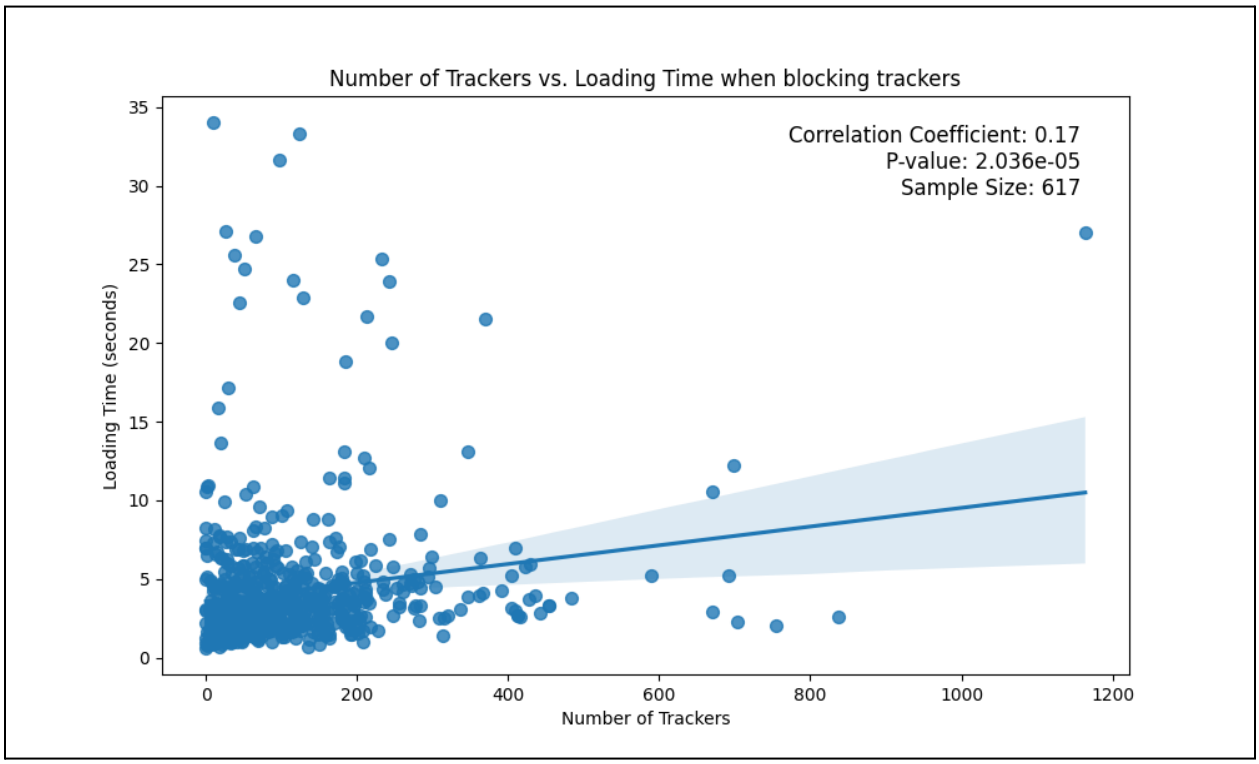

Figure 2. Graph of relationship between number of trackers versus loading time of a website when trackers are blocked with Ghostery extension

At first glance it seems trackers play an insignificant role in website loading time overall, but this is not the case. Although the two datasets by themselves do not indicate a significant relationship, when we look at each website's loading time with and without trackers blocked by Ghostery, a difference emerges that provides evidence for trackers having a tangible impact on

website loading time. Of the 576 websites that both crawls got data on, blocking trackers on the sites resulted in an average decrease in loading time of 0.469 seconds. The comparison of the loading times can be seen Figure 3 below.

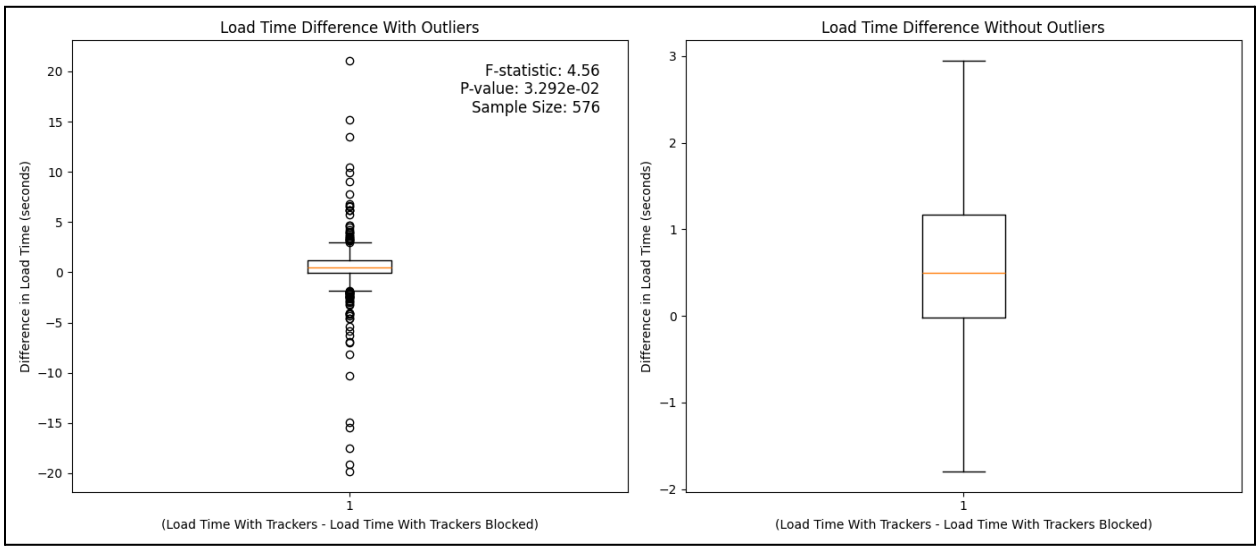

Figure 3. Comparison of loading time for each website with and without trackers blocked by Ghostery. Outliers are left out in the graph on the right side for visual purposes.

Such a large average difference across 576 websites indicates that trackers make a significant impact in an individual website's loading time (as evidenced by the low P-value of 0.0329), even if having more trackers does not strongly indicate longer loading times of websites in general. Therefore, while trackers may not be among the most prominent factors in determining website load time (in comparison to page size, for example), when controlling for each individual website's characteristics, we discovered that trackers take up a noticeable portion of the website's loading time.

Cookies and Website Loading Time. Besides looking at trackers in general, we wanted to determine if cookies, perhaps the most prominent type of tracker in terms of size, had any pronounced effect on a website's loading time. To this end, we examined the relationship of both the number of cookies on a website and its loading time, as well as the total size of cookies on a website and its loading time. Our results are presented in Figures 4 and 5 below.

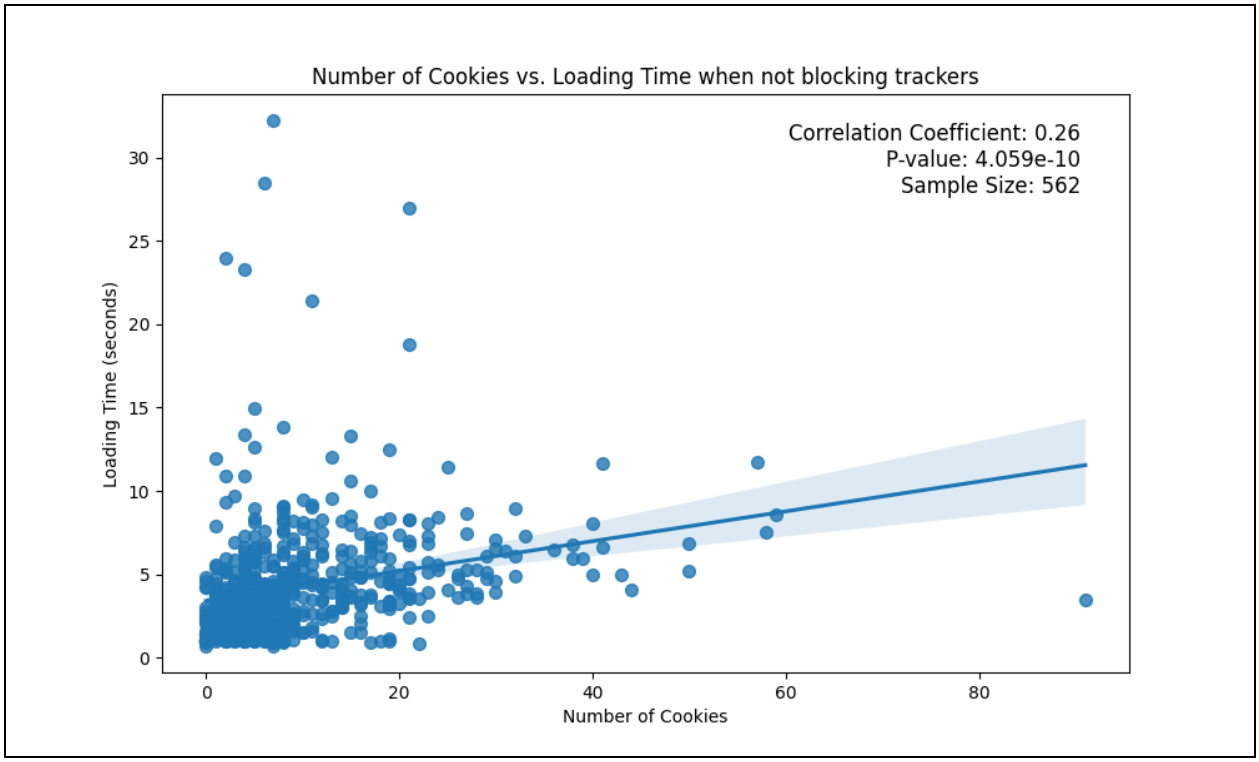

Figure 4. Graph of relationship between number of cookies and loading time of websites.

Much like with the dataset comparing the number of trackers to website loading time, we see several points that are quite distant from the main bulk of data, mostly in the response variable (loading time) axis, which likely influenced our results. The remaining cloud of data, though, follows a stronger linear trend between number/size of cookies and website loading time, as evidenced by a higher correlation coefficient of 0.26 for the number of cookies, 0.28 for total size of cookies, and 0.12 between number of trackers and website loading time, all with

extremely low P-values. Such a result matches our expectations, since we anticipated cookies to have a significant and pronounced effect on the page load time given their size and that they must be either created or retrieved when the page is loaded.

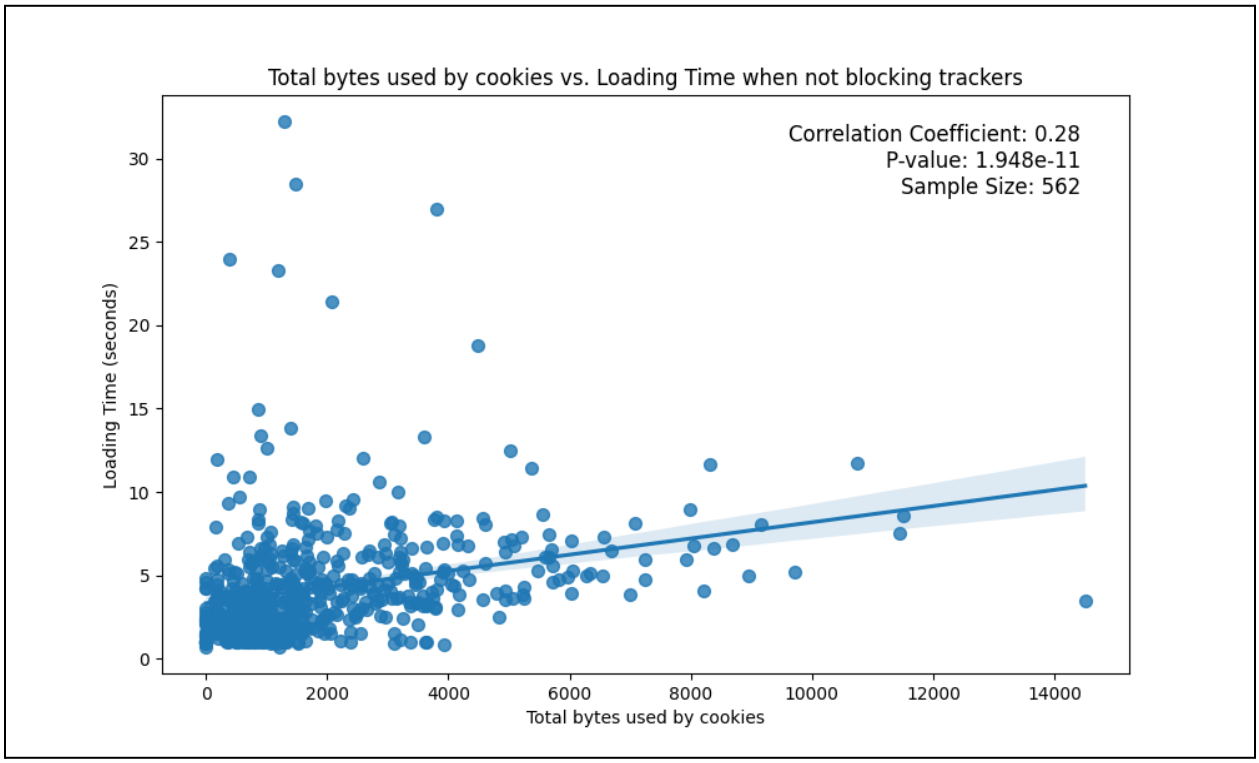

Figure 5. Graph of relationship between total size of cookies and loading time of websites

Taking into account all of these crawls, their data, and the analyses of said data, we have evidence to conclude that trackers and especially cookies do negatively impact the loading time of a website, although the extent to which they impact loading time varies heavily from site to site.

5. Conclusion

In this study, our group set out to evaluate the relationship between a website's trackers and its respective loading time by performing various web crawls to measure the amount of

tracking occurring on a site alongside its load time. Our three separate crawls, measuring number of trackers, number of cookies, total size of cookies, and load time of a site with and without trackers blocked, provide evidence that the presence of trackers (especially a large amount of cookies) on a website is associated with longer load times. While the impact of trackers on site loading time may not be as pronounced in our data as one might expect, it is important to remember the limitations of this study and the improvements that could be made to it. Further research on this subject would control the extraneous variables which could have impacted loading time by creating a controlled environment where all conditions are as uniform as possible apart from the trackers, cookies, and cookie sizes. Nevertheless, with the information that we were able to glean from this study, we demonstrate the impact trackers effect on users' web-browsing experience, and provide a basis on which more inquiries can be made.