The Tracker Tax: the impact of third-party trackers on website speed in the United States

Molly Hanson molly@ghostery.com
Patrick Lawler patrick@ghostery.com
Sam Macbeth sam@cliqz.com

May 2018

1 Introduction

Within the current Internet ecosystem, websites often include trackers – pieces of code that monitor user navigation patterns and collect personal information. Various existing studies assess the prevalence, motives and privacy concerns behind third-party online trackers (hereafter trackers). For instance, Metwalley et. al [1] demonstrate the ubiquitous nature of online tracking, explain that there are various business models for trackers, and describe how websites are "attracted by the chance to monetize visits". Englehardt and Narayanan [2] also illustrate the pervasiveness of tracking and notice that news sites tend to have the most trackers, explaining that with a lack of external funding, news sites "are pressured to monetize page views with significantly more advertising." Yu et. al [3] emphasize the privacy implications of trackers and describe how tracking companies "have the ability to collect individual users' browsing habits...across the whole Web".

While these and other previous studies tend to focus on trackers from a privacy perspective, there is a lack of existing research related to the implications of trackers on website performance. Namely, do online trackers affect page performance and are users experiencing longer page load times due to the presence of trackers? This study seeks to answer just that, by determining whether a relationship between trackers and page load speeds exists, and if so, quantifying this impact on website performance into a so-called *tracker tax*.

The need to shed light on this relationship becomes increasingly important with the recent repeal of net neutrality in the United States, as Internet service providers (ISPs) may begin to segregate the Internet into fast and slow lanes. In such a world with different Internet lanes, knowing what other factors impede website performance and contribute to longer page load times is a necessary step toward improving citizens' digital experience.

2 Data Overview and Cleaning

To assess the relationship between trackers and page load speeds, also referred to as page latency, in the desktop environment, we employed a custom-built web crawler to collect the number of trackers and page load times for the top

500 websites in the United States, as determined by Alexa [4]. The crawler was built with Selenium running Chrome, making GET requests from a server based in New York City, and using Ghostery [5] to collect:

- Count of third-party trackers: Ghostery detects third-party trackers by comparing a HTTP request with an instance in their database, which currently contains over 3,000 company trackers and 4,700 tracker patterns.

- Seconds to load page: delta between `domContentLoadedEventStart` and `requestStart`, based on the `Window.performance` API [6].

The crawler ran five times on on each domain to account for the variability that users may experience even when loading the same page. Data cleaning included removing domains with fewer than five successful measurements and excluding four Chinese websites from the sample. We excluded these websites because we suspect there are China-based trackers that are not yet accounted for in the Ghostery database. Additionally, page load times for these four websites may potentially be impacted if their servers are located in China. Since distance matters for latency and the current dataset does not incorporate a distance measure, we thought it best to exclude these Chinese websites for the following analysis. Further, to account for the variation in the data, we filtered out the fastest and slowest page loads per site so the data would be less sensitive to outliers and data collection errors.

3 Results

3.1 Tracker Ecosystem

Figure 1 shows that just over 10% of the page loads in the sample are tracker-free. Conversely, nearly 90% of page loads have at least one tracker, roughly 65% have at least 10 trackers and over 20% have 50 or more trackers. These metrics reconfirm the prevalence of online tracking, and broadly align with our previous study [7] which observes that 77.4% of page loads contain trackers. Compared to our previous study, the increase in tracker dominance seen here is likely related to one or more of the following factors:

- The sample of this study is the 500 most popular websites in the United States, while our previous study analyzed 144 million page loads across more than 12 countries. As this study only considers the most popular websites and neglects the long tail of more obscure ones, it is not surprising that a larger portion of domains in this study had a tracker present. Existing research supports this proposition, as Metwalley et al. [1] observe that "the number of trackers per service tends to increase with the popularity of the service."

- The data for this study was synthetically generated using a custom web crawler on a specific list of domains. Whereas our previous study utilized

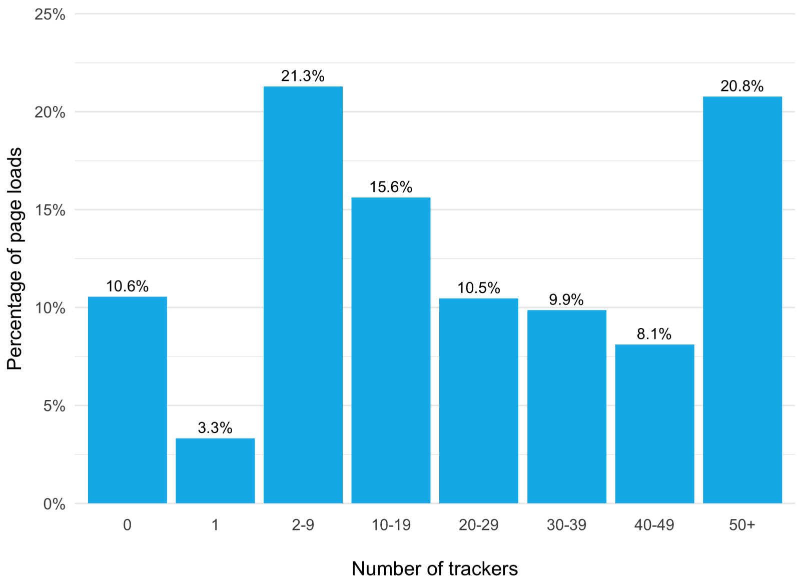

Figure 1: Distribution of the number of trackers per page load

GhostRank data, which was gathered from users of the Ghostery browser extension who had opted-in to the collection of information about trackers on pages they visit.

- Due to the use of different data sources, the definition of a *tracker* in the last study may vary slightly. Regardless, both studies validly measure and verify tracker pervasiveness throughout the web.

In any event, the focus of this study is to evaluate page performance, and to empirically quantify the effects of trackers on page load times. Due to these specific goals, a sample containing only the most popular American domains and employing a custom web crawler to scan and record tracker counts and page load times is justified.

3.2 Trackers and Page Latency

Figure 2 depicts the distribution of page load times for the sample. The data shows that only 17% of pages loaded within 5 seconds, and other than that, pages load quite slowly. It took more than 10 seconds to load nearly 60% of pages, more than 30 seconds for 18% of pages, and nearly 5% of pages took over a minute to load. This long tail cannot be ignored and suggests Internet users waste a lot of time every day simply waiting for websites to load.

Figure 2 however does not illustrate the relationship between the quantity of trackers and page latency. The number of trackers present on a website and average page load times is demonstrated in Figure 3, and unsurprisingly average page load times appear to increase with the number of trackers on a page. Note, prior to calculating average page load times by tracker count, tracker volumes with fewer than 5 observations and page latency outliers within each tracker count, identified using the interquartile range rule, were excluded.

Initially, a simple linear regression was used to model the relationship between the number of trackers on a website and the average time to load that page (Figure 4). With an adjusted-R^2 value of 0.802, this strong, positive linear model suggests that each additional tracker adds, on average, 0.5 seconds to the overall page load time. As the data points show a curved trend, a second model with a quadratic term was used to model this relationship (Figure 4). Adding the quadratic term realizes an adjusted-R^2 value of 0.836, so this more complex model explains roughly an additional 3.6% of the variation in average page loads, using the number of trackers on a page. Both the intercept and coefficient for the quadratic term are highly significant and this quadratic model suggests trackers have an increasing impact on page load times: as the number of trackers grows larger, additional trackers add even more time to page loads.

It is important to consider that linear regression has several underlying assumptions. Upon analysis of the residuals, the two models outlined above exhibit heteroscedasticity – uneven variance of the error terms – and thus violate one of the underlying assumptions of linear regression. In such instances, variance-stabilizing transformations can be applied to the data, and here the Box-Cox test was employed to determine if the response variable – average page load time – should be transformed. The results of the test suggest log transforming the response variable will realize the best-fitting model, since the log-likelihood function is maximized when λ is roughly 0.

The transformed data (after taking the natural log of the average page load times) is shown in Figure 5. The log-linear model has an adjusted-R^2 value of 0.885, and both the intercept and variable coefficients are significant. This model indicates a compounding effect: if the tracker count increases by 1, we expect the average page load time to increase by 2.5%. Lastly, the log-quadratic model in Figure 5 does not show an improvement in adjusted-R^2 over the log-linear model, and staying true to the principle of parsimony, the simpler log-linear model is preferred.

3.3 Protection from Trackers

Internet users have access to various technologies that block trackers from loading on websites. The benefits range from faster page loads and eliminated clutter to data protection and enhanced privacy. To observe the difference in page latency when trackers are allowed rather than blocked, the web crawler referenced in Section 2 was run with the same parameters; however, Ghostery [5] was enabled to block all trackers. The dataset containing page load times and volume of trackers per page with all trackers blocked underwent the same data cleaning process described in Section 2.

Figure 6 compares page latency with no trackers blocked and with all trackers

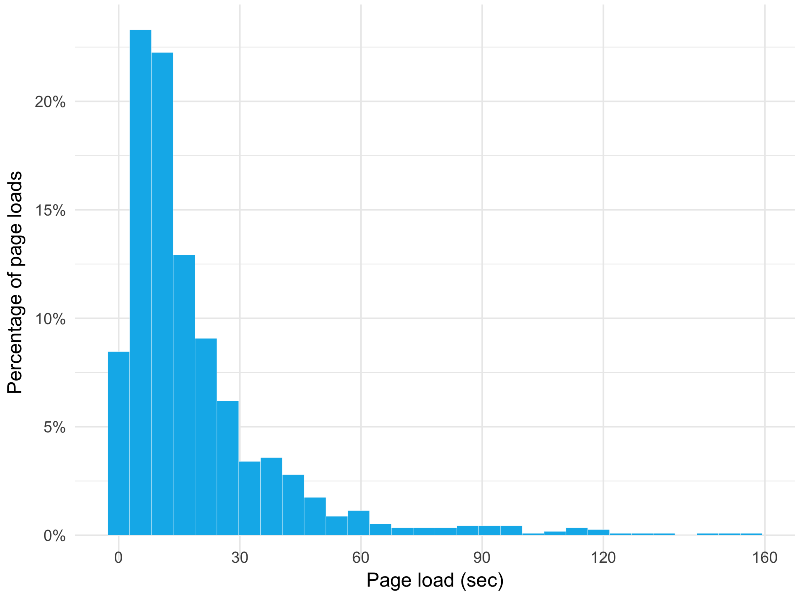

Figure 2: Distribution of page latency, in seconds

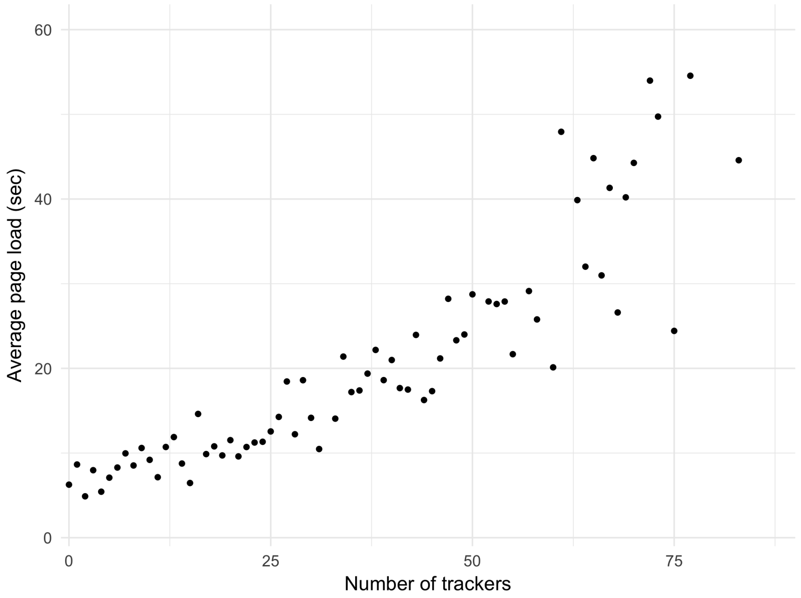

Figure 3: Relationship between number of trackers on a website and average page load times, in seconds

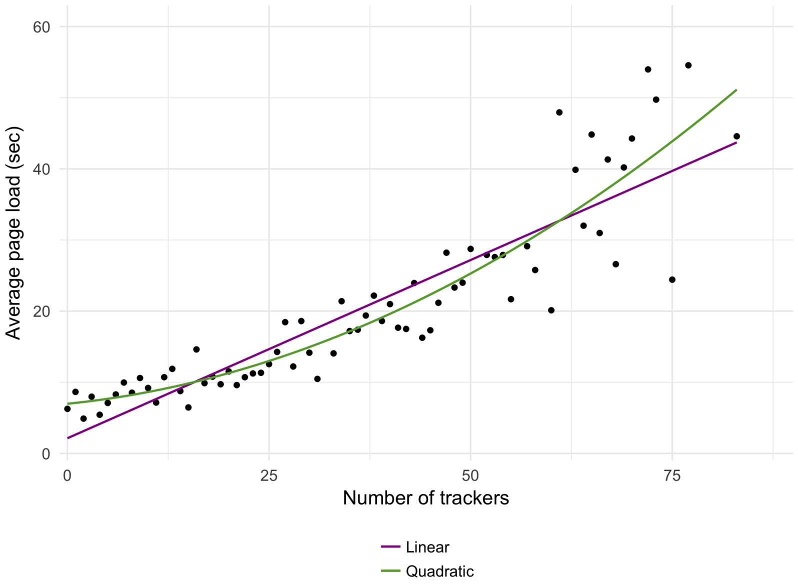

Figure 4: Linear and Quadratic Models to quantify the relationship between number of trackers on a website and average page load times, in seconds

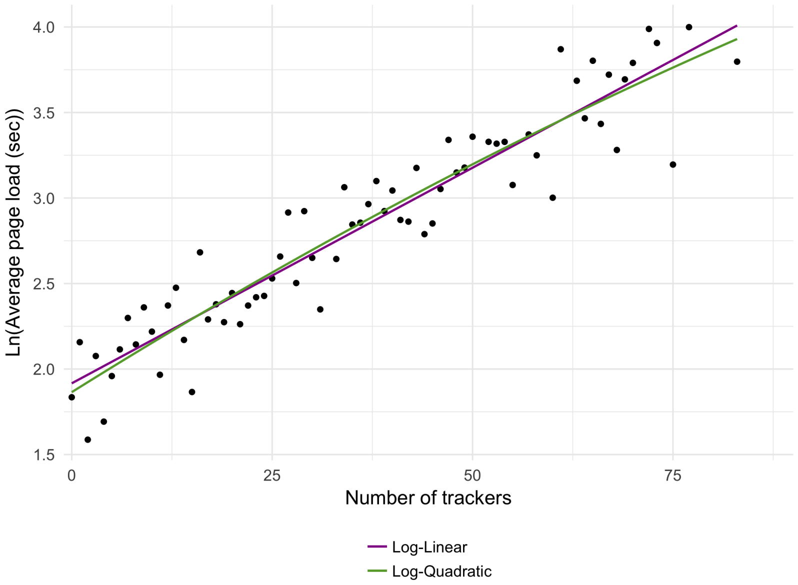

Figure 5: Linear and Quadratic Models to quantify the relationship between number of trackers on a website and natural log of the average page load times, in seconds

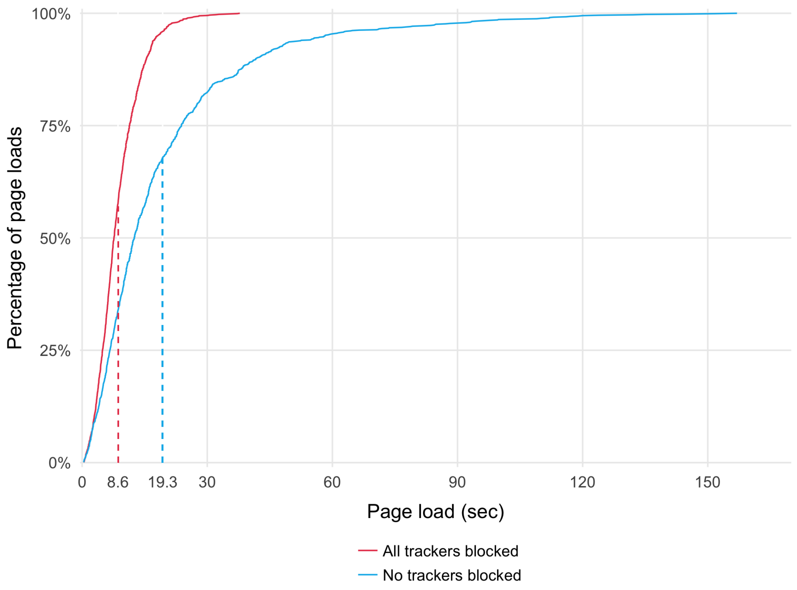

Figure 6: CDF of page load times, with all and no trackers blocked

blocked, with mean load times of 19.3 and 8.6 seconds, respectively. Thus, on average, websites take more than twice as long to load with trackers unblocked and the average page load time is increased by 10 seconds.

The time savings of using tracker blocking technologies are even more drastic when considering a subset of the 10 slowest domains in the sample. When blocking all trackers on the slowest domains, load times are on average 10 times faster and users save an average of 84 seconds per page load. These time savings by domain are represented graphically in Figure 7.

In terms of domains with the highest average volume of trackers, Figure 8 demonstrates that the domains with the most trackers have on average 10 times more trackers present when trackers are not blocked. Additionally, among these domains, on average there are 93 fewer trackers present per page load with all trackers blocked. This phenomenon, whereby there are significantly more trackers present when trackers are unblocked as compared to blocked, is due to piggybacking. Piggybacking refers to one tracker placed directly on the website giving access to other "piggybacking" trackers that are not originally on the site. This can create a snowball effect, where trackers bring in more trackers that can then bring in even more trackers, and so on. As suggested above, each additional tracker slows down a website more than previous ones, so this common occurrence has notable performance implications.

Moreover, this piggybacking practice brings up multiple privacy concerns: piggybacked trackers often include data companies that sell data to other businesses looking to target people, and since these trackers are not directly on the

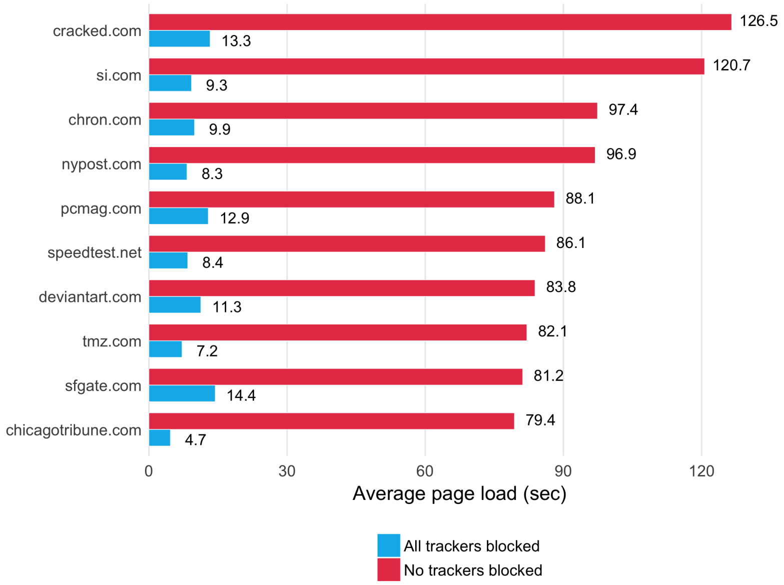

Figure 7: Average page load time of 10 slowest domains without tracker blocking, compared to the average page load time with all trackers blocked

website, site owners do not have insight into who is collecting data about their users and some may not even be aware such intrusion is occurring. Although these are noteworthy concerns, this falls outside the scope of this study.

4 Implications

Online tracking is clearly pervasive, with nearly 90% of the most popular sites in the United States having at least one third-party tracker present. In addition to the privacy concerns of online trackers, the findings presented in this study indicate a strong, positive link between the number of third-party trackers on a page and the time it takes to load that page. Generally, the more trackers on a website, the longer a user will have to wait for that site to load. Quantifying this relationship depends on the model used, however the optimal model outlined in Section 3.2 shows a compounding effect: if the tracker count increases by 1, the average page load time is expected to increase by 2.5%.

Various tracker blocking tools are available that users can not only use to protect their privacy but also speed up their browsing. On average, websites take more than twice as long to load when trackers are not blocked and the slowest sites take 10 times longer to load. These added waiting times are not trivial, especially as the population spends increasingly more time online [8].

It is important to consider that the study examined only the top US websites

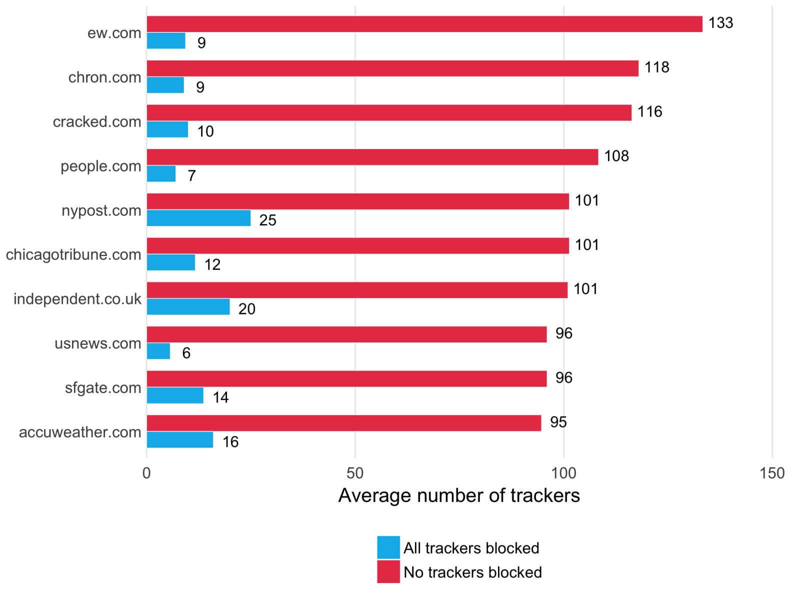

Figure 8: Average number of trackers on the 10 domains with the most trackers without tracker blocking, compared to the average tracker count with all trackers blocked

accessed locally. Potential future research includes expanding this analysis to other regions, to determine if similar trends exist outside the United States. Even though other nations may not be experiencing the same shift in their local ISP industry, the tracker tax itself is real. Regardless of policy, Internet users across the globe who may not fully grasp the viral pervasiveness of online tracking, are also likely to be unaware of the associated time costs.

Further research may also include measuring additional performance implications of trackers. For instance, some trackers make requests to other servers and data is transferred in the process – and this data transfer bears real monetary costs to the user. The current study is limited to the desktop environment, however researching the relationship between trackers and data volume transferred may be better suited for a mobile study. In expanding this study to mobile, the data volume that is consumed by trackers could be translated to the out of pocket expense suffered by the user, in addition to the more subjective dollar value of the user's wasted time expended while waiting for pages to load.

Existing research also suggests a relationship between increased page load times and increased "bounce rates", the frequency at which site visitors navigate away from or abandon a page. A recent study by Akamai [9] found that "a two-second delay in load time hurt bounce rates by up to 103%". Trackers therefore are a "cost" to the website owner: users are more likely to leave websites the longer they take to load [10], and trackers contribute to these longer load times. Still, assuming rational behavior on the part of website owners, there must be a compensating value or benefit of adding trackers to their site, that outweighs an increased bounce rate. Given the wide variety of tracker use cases, the net value to the site owner from using trackers on their website may come in the form of insight into visitor behaviors, advertising dollars or improved interaction with customers. Moreover, longer page load times on news and e-commerce sites have been linked with lost revenue [11, 12]. Therefore, to justify the use of trackers, site owners must believe that the knowledge, learnings and ad revenue associated with trackers on their site offset this lost revenue due to trackers slowing down the page.

Future research into bounce rates and page load speeds could be used to calculate a hypothetical tracker value measure: given the additional time trackers add to page loads, and that slower pages lead to a loss in site traffic, one tracker should provide the same value as this lost site traffic. Note that this measure would assume trackers are being added rationally by the website owners, which is not the case as observed in the above Section 3.3 that explains how trackers give access to piggybacked trackers which are not placed on the site initially. Given that this study focuses on page load times for the most popular homepages in the United States, and that domain category, function and funnel page are also likely to be notable factors which also influence bounce rates, a future analysis may include studying how trackers on e-commerce sites affect the likelihood for bounce rates and thus lost revenue.

5 Summary

Indeed, this study, which was conceptualized when net neutrality was law, shows that trackers have a pervasive, negative impact on page load times. Now in a world without net neutrality regulations, users and their browsing speeds may be squeezed from both sides – by the ISP and the online tracker ecosystem. Thus, we may start to see more of a two prong tracker tax: the direct monetary impact imposed by the ISP and the more subjective dollar value cost to the user who has longer load times and therefore longer, unproductive ‘dead time’ imposed by trackers.

Perhaps now more than ever, in the wake of the net neutrality repeal, users must consider the performance implications of browsing online without protection from trackers. And, through the implementation of tracker blocking technologies, users can not only protect their privacy, but also speed up their browsing experience by avoiding the tracker tax.

References

[1] H. Metwalley, S. Traverso, and M. Mellia, “Using passive measurements to demystify online trackers,” *Computer*, vol. 49, pp. 50–55, Mar 2016.

[2] S. Englehardt and A. Narayanan, “Online tracking: A 1-million-site measurement and analysis,” 2016.

[3] Z. Yu, S. Macbeth, K. Modi, and J. M. Pujol, “Tracking the trackers,” in *Proceedings of the 25th International Conference on World Wide Web*, pp. 121–132, International World Wide Web Conferences Steering Committee, 2016.

[4] Alexa Internet, “Alexa Internet.” https://www.alexa.com/topsites/countries/US. Accessed: 2017-11-16.

[5] Ghostery, “Ghostery.” https://ghostery.com/.

[6] MDN Web Docs, “Window.performance.” https://developer.mozilla.org/en-US/docs/Web/API/Window/performance, August 15, 2017.

[7] S. Macbeth, “Tracking the Trackers: Analysing the global tracking landscape with GhostRank,” July 2017.

[8] Nielsen, “The Nielsen Total Audience Report: Q1 2017.” http://www.nielsen.com/us/en/insights/reports/2017/the-nielsen-total-audience-report-q1-2017.html, July 12, 2017.

[9] Akamai, “The state of online retail performance.” https://www.soasta.com/wp-content/uploads/2017/04/State-of-Online-Retail-Performance-Spring-2017.pdf, Spring 2017.

[10] R. Elliott, “How Page Load Time Affects Bounce Rate and Page Views.” https://www.section.io/blog/page-load-time-bounce-rate, August 10, 2017.

[11] M. Chadburn and G. Lahav, “A faster FT.com.” http://engineroom.ft.com/2016/04/04/a-faster-ft-com/, April 4, 2016.

[12] J. Bixby, “Slow Shopping Cart Pages Are Killing Conversions. An Optimized Page Got a 66% Conversion Lift!.” https://unbounce.com/conversion-rate-optimization/case-study-your-slow-shopping-cart-pages-are-killing-conversions-heres-what-you-can-do-about-it, November 29, 2011.