Deconstructing Google's Web Light Service

Conference Paper · April 2020

DOI: 10.1145/3366423.3380168

Ammar Tahir
Lahore University of Management Sciences
1 PUBLICATION **3** CITATIONS

Deconstructing Google's Web Light Service

Ammar Tahir, Muhammad Tahir Munir, Shaiq Munir Malik, Zafar Ayyub Qazi, Ihsan Ayyub Qazi
Department of Computer Science, SBASSE, LUMS, Pakistan

ABSTRACT

Web Light is a transcoding service introduced by Google to show lighter and faster webpages to users searching on slow mobile clients. The service detects slow clients (e.g., users on 2G) and tries to convert webpages on the fly into a version optimized for these clients. Web Light claims to significantly reduce page load times, save user data, and substantially increase traffic to such webpages. However, there are several concerns around this service, including, its effectiveness in, preserving relevant content on a page, showing third-party advertisements, improving user performance as well as privacy concerns for users and publishers.

In this paper, we perform the first independent, empirical analysis of Google's Web Light service to shed light on these concerns. Through a combination of experiments with thousands of real Web Light pages as well as controlled experiments with synthetic Web Light pages, we (i) deconstruct how Web Light modifies webpages, (ii) investigate how ads are shown on Web Light and which ad networks are supported, (iii) measure and compare Web Light's page load performance, (iv) discuss privacy concerns for users and publishers and (v) investigate the potential use of Web Light as a censorship circumvention tool.

CCS CONCEPTS

• **Information systems → World Wide Web**.

KEYWORDS

Web Light, transcoding, page load, privacy, censorship, circumvention

ACM Reference Format:
Ammar Tahir, Muhammad Tahir Munir, Shaiq Munir Malik, Zafar Ayyub Qazi, Ihsan Ayyub Qazi. 2020. Deconstructing Google's Web Light Service. In *Proceedings of The Web Conference 2020 (WWW '20), April 20–24, 2020, Taipei, Taiwan.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3366423.3380168

1 INTRODUCTION

Mobile devices have become the predominant mode of Internet access; since October 2016 more websites have been loaded on mobile devices than on desktop computers [14]. As a result, mobile subscriptions have experienced rapid growth, with 7.9 billion subscriptions at the end of 2018 [12]. Developing countries, in particular, are experiencing explosive growth, with more than 98% mobile phone adoption [13]. At the same time, the complexity of websites has been increasing; since April 2016, the median mobile page size has increased by 86% resulting in a median mobile webpage taking 18.7 s, 300% more time than the median desktop page [5].

To improve the web browsing experience on slow mobile clients, Google in 2015 introduced a transcoding service, Web Light, which coverts webpages on the fly into pruned versions, optimized for such clients. The service is operational in select countries like India, Brazil, Indonesia, and Pakistan[1]. Slow mobile clients searching from the Chrome browser, Android browser or Google Go [17] are automatically shown results from Web Light pages. Google first detects users on a slow network (e.g. 2G), and automatically redirects them to the Web Light service. Some publishers report seeing a significant amount of traffic coming from the Web Light service [18, 20]. Google says the service can help in (i) improving user perceived performance, (ii) saving user data, and (iii) increasing traffic to webpages, which they expect will help monetize publishers [11].

However, despite being around for a few years, Web Light remains shrouded in mystery. There is little information provided on how Web Light works; *how* and *when* Web Light transcodes webpages? Which web objects (e.g., JavaScript, images, videos) does it transcode? How are advertisements on a webpage handled? At the same time, there are growing concerns from both publishers and users, including: Web Light breaking webpages[2] [18, 25], users being directed to Web Light pages by default on their mobile phones [18], and Web Light taking away the advertisement revenue of the publishers as it supports only a small number of advertisements [19].

In this paper, we perform the first independent, empirical analysis of Web Light to shed light on these concerns and provide users and publishers insights on making informed choices. Through a combination of actual webpages and synthetic webpages, we deconstruct and analyze Web Light service. Specifically in this study, we (i) examine how Web Light modifies webpages, (ii) investigate how advertisements are shown on Web Light pages and which ad networks are supported, and (iii) quantify improvements in page load performance. Our study also highlights important privacy concerns; lack of consent from publishers and users, as well as concerns around Web Light's proxy architecture. We also discover that Web Light can provide fast access to censored content, providing incentive for users to employ it as a censorship circumvention tool.

Below, we highlight the key insights from our experimental evaluation and analysis of Google's Web Light service. Table 1 provides a summary of these insights.

- **Complexity of Web Light pages:** Web Light transcodes webpages by either removing images, reducing image resolutions, removing JavaScript, changing CSS, and/or replacing

WWW '20, April 20–24, 2020, Taipei, Taiwan

ACM ISBN 978-1-4503-7023-3/20/04.
https://doi.org/10.1145/3366423.3380168

[1] These alone are among the ten most populous countries in the world and cumulatively account for over 2 billion of the world's population [32].

[2] Typical concerns include Web Light breaking or not displaying clearly webpage menus, social media icons, shopping carts and embedded videos.

Key Insights	Description
(1) Web Light Transcoding	Web Light transcodes webpages by either removing images, reducing image resolutions, removing JavaScript, changing CSS, and/or replacing videos with an image.
(2) Web Transcoding Details	Web Light reduced the median page size by 12x. Transcoded pages contained 15x less images and 3.5x less scripts in the median case.
(3) Impact on Advertisements	Web Light shows only up to three ads on any page and does not support many popular advertising networks. Among the 26 ad networks we observe in our dataset, Web Light only supported 9. This can significantly limit websites' ad revenues.
(4) Page Load Performance	Web Light reduced the median PLT by 5.4x. Both transcoding and caching of webpages at the Web Light proxy contributes to a reduction in PLTs.
(5) Web Privacy	Web Light's proxy architecture supports HTTPS pages but does not preserve end-to-end encryption. Thus, Web Light can potentially build users' browsing profiles and read website content being viewed by users.
(6) Censorship Circumvention	Web Light provides fast access to censored content and significantly outperforms Tor (e.g., median PLT improvement ranged from 6.8x-24x) and the Hotspot Shield VPN (e.g., median PLT improvement ranged from 2.4x-10.1x). This can serve as an incentive for users to employ Web Light as a censorship circumvention tool.

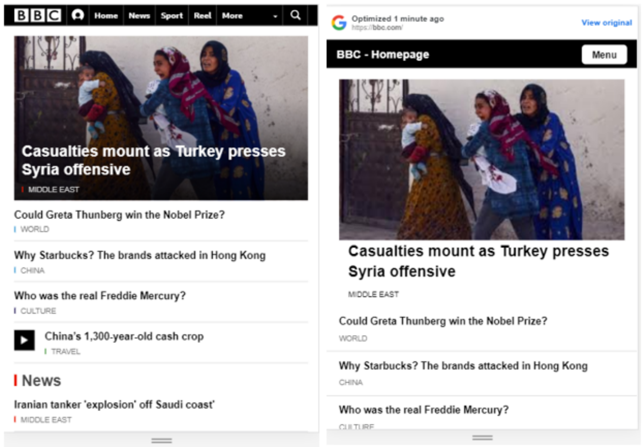
Table 1: Key insights from our experimental evaluation and analysis of Google's Web Light service.

videos with an image. Our measurements on top 5000 Alexa pages shows that Web Light reduces the median page size by 12×, transcoded pages contain 15× less images and 3.5× less scripts in the median case (§4).

- **Impact on Advertisements:** Web Light shows only up to three ads on any page and does not support many popular advertising networks. For example, among the 26 ad networks we observe in our dataset, Web Light only supported ads from 9 ad networks. This can significantly limit websites' ad revenues (§5).
- **Page Load Performance:** Our experiments show that Web Light improves page load performance; reducing the median Page Load Time (PLT) by 5.4×. Both transcoding and caching of webpages contribute to a reduction in PLTs (§6).
- **Web Privacy Concerns:** Web Light's proxy architecture supports HTTPS pages but does not preserve end-to-end encryption. Thus, Web Light can potentially build users' browsing profiles and read website content being viewed by users. We discuss Web privacy concerns in §7.
- **Censorship Circumvention Tool:** Through our experiments, we find that Web Light provides fast access to censored content and significantly outperforms Tor [23] and the Hotspot Shield VPN [7] in terms of PLTs (§7). This can serve as an *incentive* for users to employ Web Light as a censorship circumvention tool.

Our analysis reveals that even though Web Light can provide significant improvements in page load performance and reduction in data usage, there are legitimate concerns around privacy, ad revenue and censorship circumvention. We hope this study brings to light these tradeoffs and helps users and publishers to make informed choices about Web Light. Whether the potential benefits outweigh the potential harms or otherwise is for users, publishers and other stakeholders to decide.

The remainder of this paper is organized as follows: we first provide an overview of Web Light (§2), then describe our experimental methodology to deconstruct Web Light (§3). We discuss our experimental results and analysis as follows: complexity of Web Light pages (§4), impact on advertisements (§5), impact on page load performance (§6), and privacy concerns and the use of Web Light as an Internet censorship circumvention tool (§7). We finally discuss related work (§8) and the broad implications of using Web Light service (§9).



Casualties mount as Turkey presses Syria offensive
MIDDLE EAST

Could Greta Thunberg win the Nobel Prize?
WORLD

Why Starbucks? The brands attacked in Hong Kong
CHINA

Who was the real Freddie Mercury?
CULTURE

China's 1,300-year-old cash crop
TRAVEL

News

Iranian tanker 'explosion' off Saudi coast'
MIDDLE EAST

Casualties mount as Turkey presses Syria offensive
MIDDLE EAST

Could Greta Thunberg win the Nobel Prize?
WORLD

Why Starbucks? The brands attacked in Hong Kong
CHINA

Who was the real Freddie Mercury?
CULTURE

Figure 1: The www.bbc.com homepage (left) and its Web Light transcoded version (right).

2 WEB LIGHT OVERVIEW

In this section, we provide an overview of Google's Web Light service. In particular, we discuss how the service is architected, and how users access the service.

2.1 Web Light Architecture and Usage

The Web Light service is architected as a HTTPS proxy service that transcodes webpages into lighter pages so that they load faster on slow clients while saving data. While the official page of Web Light [11] says, "*Web Light pages preserve a majority of the relevant content*", it does not describe *how* and *when* Web Light transcodes pages and *which* objects does it transcode and which it does not. Through our experimental evaluation, we find that it transcodes webpages by either removing images, reducing image resolutions, removing JavaScript, and/or changing CSS and it does not provide support for cookies. Webpages are currently transcoded for searches from the Chrome browser and the Android browser (version 2.3+), as well as Google Go [17]. Users on a slow network (e.g., 2G) are automatically directed to the search results from Web Light service. Alternatively, users can also use the Web Light URL to use the service (e.g., https://googleweblight.com/i?u=https://bbc.com). The Web Light service is offered in select countries including India, Indonesia, Brazil, and Pakistan.

Web Light does not transcode *all* webpages, such as websites that require users to login before using them (e.g., www.amazon.com) and some video streaming websites (e.g., www.youtube.com) because Web Light does not transcode videos into lower resolution videos[3]. To keep the sizes of webpages small, Web Light limits the number of ads shown on a page to three. The three ads are chosen in the *order* in which they are requested by the original page. Moreover, as of now Web Light only supports a subset of the ad networks. If publishers do not want their pages to be transcoded, they can explicitly opt out by setting the HTTP header "Cache-Control: no-transform" in their page response. If Web Light sees this header, the requested page will not be transcoded.

We find that the Web Light service uses at least two proxy servers to deliver a webpage. To test this we created a synthetic webpage, hosted on a server controlled by us. We observed that the IP address of the Web Light server the client was interacting with was *different* from the one our server was interacting with.

3 METHODOLOGY

In this section, we describe the experimental methodology we devised to deconstruct how Web Light transcodes webpages and understand its impact on mobile web performance. We use a combination of actual webpages and synthetic webpages that we construct to systematically understand *how* and *when* Web Light transcodes webpages and how it affects user perceived performance.

We access Web Light from the city of Lahore in Pakistan (where the service is operational) by using the Web Light URL from a mobile device. We use Selenium [8] with Google ChromeDriver [22] to obtain the complete source code of a webpage. ChromeDriver is a standalone server that implements the W3C WebDriver standard.

3.1 Collecting Web Light Pages

Our dataset comprises the landing pages of Alexa top 5000 websites as well as their Web Light versions. The dataset contains the base HTML of the original (non-transcoded) webpages and the Web Light transcoded pages along with all external objects fetched by a page. We found that Web Light transcoded only 3440 websites from the list of Alexa top 5000 websites. If the Web Light proxy cannot transcode a page, it redirects it to the original website (via the direct path that does not use the proxy). Thus, we remove all such pages from our dataset. Using this dataset, we compare sizes of the base HTML, images, and scripts of the original webpages with their Web Light versions.

3.2 Creating Synthetic Webpages

To get a deeper insight into how and when Web Light transcodes different types of webpages, to to answer questions such as, *does Web Light transcode pages that have a small size to begin with*? *does it serve lower resolution images only if an image is above a certain resolution threshold? if yes, then what is that threshold?*, we construct different kinds of synthetic pages and host them on a server under our control and observe how and when Web Light transcodes them. With synthetic pages, we measure thresholds at which Web Light transcodes a page (instead of redirecting to original page) across different objects such as images and JavaScript. To this end, we vary image sizes from 4 KB to 82 MB and also vary the number of images from 1 to 100. To see how Web Light manages scripts, we embed inline scripts as well as external JavaScript in our test pages. CSS files of different sizes and in-line CSS are also embedded and tested. Finally, we also use synthetic pages to determine if Web Light uses any other intermediate proxy service (e.g., either within a Google datacenter or across datacenters) for resilience, security, or other purposes.

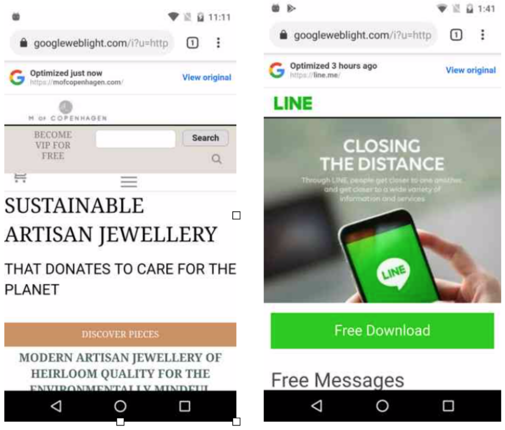

googleweblight.com/i?u=http
Optimized just now
https://mofcopenhagen.com/
View original
M of COPENHAGEN
BECOME VIP FOR FREE
Search
SUSTAINABLE ARTISAN JEWELLERY
THAT DONATES TO CARE FOR THE PLANET
DISCOVER PIECES
MODERN ARTISAN JEWELLERY OF HEIRLOOM QUALITY FOR THE ENVIRONMENTALLY MINDFUL
googleweblight.com/i?u=http
Optimized 3 hours ago
https://line.me/
View original
LINE
CLOSING THE DISTANCE
Free Download
Free Messages

Figure 2: Freshly transcoded page (on the left) vs cached Web page (on the right)

3.3 Measuring Page Load Performance

To measure user-perceived page load performance, we use a private instance of WebPageTest [10] and a Nexus 5 phone – a device with 2 GB RAM and 4 CPU cores each with a maximum frequency of 2.26 GHz – along with a 3G connection[4]. We fetch the landing pages of 100 randomly selected websites (that were being transcoded by Web Light) from Alexa top 1000 websites and collect several page load performance metrics (discussed below). We repeat our experiment five times, for both Web Light transcoded pages as well as the original webpages.

Next, to analyze the impact of caching at the Web Light proxy, we again fetch the same 100 randomly selected webpages after a span of two days (to reduce the likelihood of finding a cached page) but only repeat the experiment twice; one run to obtain the *non-cached* version of the page and the other to obtain the *cached* version. We classify the Web Light transcoded pages into two categories: *cached pages* and *freshly transcoded pages* (i.e., the non-cached page). This information is available in the top banner of each Web Light transcoded webpage. The top banner of a freshly transcoded page reads "Optimized Just now" whereas in case of a cached page, it shows how long ago was the page transcoded and cached as shown in Figure 2.

QoE metrics. To study the impact of Web Light on mobile QoE, we measure three commonly used metrics: Page Load Time (PLT), Time to First Byte (TTFB), and the Speed Index (SI) [9]. PLT is the time until all objects on a page have been loaded (i.e., when the onload event fires), while TTFB is the time elapsed since the request is sent and the first byte of the payload is received at the client. SI is the average time at which *visible* parts of the page are displayed.

[3] However, Web Light does replace videos with an image on some websites.

[4] We tested Web Light over different cellular connection types (e.g., EDGE, 3G and 4G) and found no evidence for differential transcoding based on the connection type.

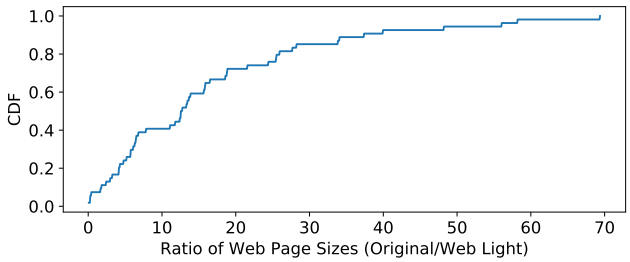


Figure 3: Ratio of page sizes (i.e., base HTML size plus the size of all fetched objects) of original webpages and their Web Light versions.

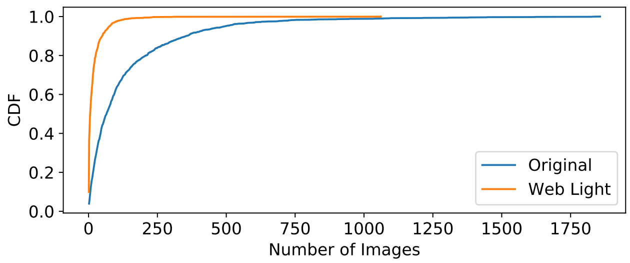


Figure 4: Comparison between the number of images in the original webpages and their Web Light versions.

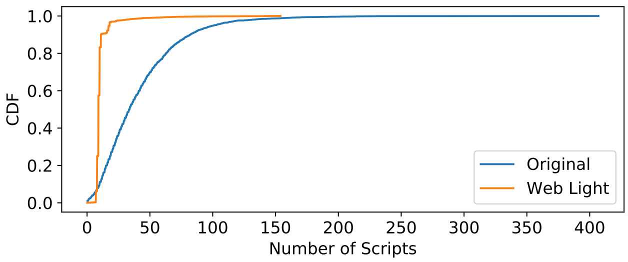


Figure 5: A comparison between the number of scripts in the original webpages and their Web Light versions.

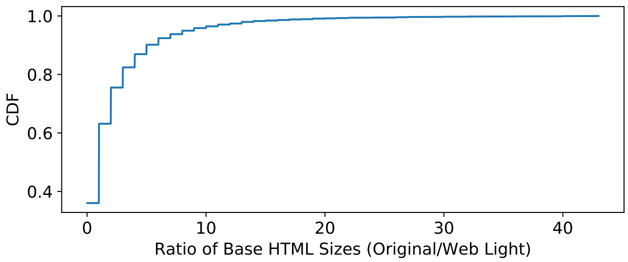


Figure 6: Ratio of base HTML sizes of the original webpages and their Web Light versions.

4 COMPLEXITY OF WEB LIGHT PAGES

To understand how Web Light pages differ from regular webpages, we analyze our dataset that contains the original and the Web Light transcoded versions of approximately 5000 webpages. Based on our analysis, we make the following observations.

(a) *Web Light reduces the median page size by 12× compared to regular pages.* The median page sizes for the original and transcoded pages were 2 MB and 170 KB, respectively. Figure 3 shows the ratio of page sizes of regular webpages and their Web Light versions. Observe that Web Light reduces the median page size by 12× whereas the maximum size reduction in page size was 68×.

(b) *Web Light pages contain 15× less images than regular pages in the median case.* We find that for regular pages, the median webpage contains 60 images compared to only 4 in case of Web Light pages as shown in Figure 4. Moreover, all the images served by Web Light are encoded in base64 format (thus part of the base HTML) with a *reduced* resolution. Our experiments show that one of the most frequently occurring base64 encoded image resolution was 300x180. Using base64 encoding for images increases the size of the base HTML page but can reduce additional round trips needed for separately downloading images, which can be particularly beneficial on slow networks.

(c) *Web Light pages contain 3.5× less number of scripts than regular pages in the median case.* We find that Web Light proactively removes scripts. In particular, we find that the median Web Light page contains 9 scripts whereas the original page has 32 scripts as shown in Figure 5. Interestingly, for 5% of the webpages the Web Light version has *more* scripts than their counterparts because Web Light adds at least eight scripts to each transcoded page. Some of these scripts are responsible for displaying custom objects (e.g., the top banner which tells *when* the was page was last transcoded). Some scripts handle exceptions (e.g., if we click on a button that was originally based on JS, these scripts trigger a pop up informing the user that these fields are disabled). One script renders cover image while another one converts all the *href* links on a page into Web Light links by appending googleweblight.com to them. Moreover, there can be only three scripts for fetching ads in a Web Light transcoded page.

(d) *Web Light increases the size of the base HTML for about 33% of the webpages.* Figure 6 shows the ratio of sizes of the base HTML of regular and Web Light pages. Observe that Web Light reduces the size of the base HTML for majority of webpages. However, 33% of transcoded pages have larger base HTML sizes compared to regular pages. This generally happens for webpages that have relatively small sizes and thus transcoding provides little or no reduction in their size. However, the additional scripts added by Web Light to the base HTML and the encoding of all first party images in base64 format leads to a net increase in size. While this inflates the time to download the base HTML file, it can potentially reduce the number of round trips.

(e) *Web Light replaces videos with images on several webpages and can also remove embedded videos.* In order to understand how Web Light deals with videos, we conduct experiments on 30 video streaming sites. We find that 20 out of these 30 webpages are transcoded whereas 10 are not transcoded. On the 20 transcoded pages, we observe the following: (i) 18 webpages had their videos converted to images (e.g., https://www.tune.pk and https://www.metacafe.com/), (ii) one webpage originally allowed users to download a video or play it directly but Web Light did not allow the video to be played directly. However, it did allow the video to be downloaded, and (iii) one webpage had a playable video on the transcoded page but the video quality was unchanged. Interestingly, we find that Web Light does not transcode some video streaming websites (e.g., https://www.youtube.com/, https://www.dailymotion.com/pk, and https://www.veoh.com/) and redirects users to the regular page. It is not clear why Web Light transcodes some video streaming websites but not others. The lack of clarity can make it difficult for publishers to adapt their webpages to attract more users to their sites.

4.1 Controlled Experiments

To take a deep dive into *how* and *when* Web Light transcodes pages, we systematically construct several synthetic pages, host them on a server controlled by us and fetch them via Web Light. Below, we describe the key insights from these experiments:

- Web Light attempts to transcode a text-only webpage *only* when the size of the text in the base HTML page is greater than 150 bytes. When transcoding a page, it embeds its own scripts and some other elements (e.g., new `<div>` tags), which are between 10 KB-12 KB in size. It does not modify the webpage text.
- Images are transcoded *only* when the webpage also contains text greater than 150 bytes. We experimented with webpages containing 1-100 images with different image resolutions and sizes (4 KB to 84 MB), but images alone never trigger transcoding by Web Light.
- When images in a webpage are transcoded (i.e., in the presence of text), their width is reduced to 300 pixels (if the width of the original image is greater than 300) and their height is calculated in proportion to the height of the original image. In other cases, images are not transcoded.
- By default, Web Light removes all scripts except scripts within iframes that are used for serving ads. It places publisher's ads back into the Web Light page (after optimizing the content) by either directly including ad tags (in case of mobile-friendly pages) or modify the ad tag's parameters to request a mobile ad from the appropriate ad network (in case pages are not mobile-friendly), colors remain same but font and other elements such as button and div shapes change.
- We find that Web Light removes all images fetched through a JS (except ads served via scripts within an iframe).
- All CSS, either external or internal (i.e., using a `<style>` element in the `<head>` section), are converted to inline CSS (i.e., CSS that uses the style attribute in HTML elements) by Web Light. Web Light does not retain all the attributes of the original CSS; it retains the color of the HTML object but may change its shape (e.g., round input field to rectangular). This indicates that it perhaps maintains some pre-defined list of attributes when transcoding.

5 IMPACT ON ADVERTISEMENTS

To keep the size of webpages small, Web Light limits the number of ads shown on a page to three. The three ads are chosen in the *order* in which they are requested by the original page. This is achieved as follows: once the Web page is transcoded, Web Light places the publisher's ads back into the Web Light page. This is done by detecting the existing ads on a publisher's page. If the original page is mobile-friendly, the ad tags are included directly. If the original page is not mobile-friendly, Web Light modifies the ad tag's parameters to request a mobile ad from the appropriate ad network.

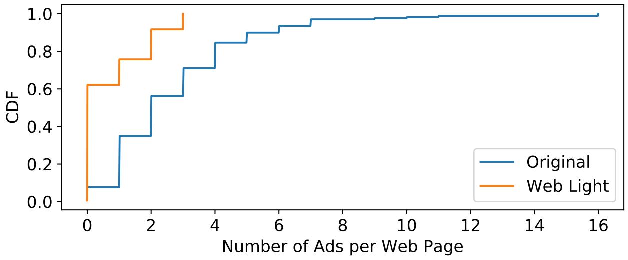


Figure 7: Distribution of the number of ads in the original and Web Light versions of webpages.

As of now, Web Light only supports a subset of the ad networks such as eSovrn, Zedo, AdSense, and Google Publisher Tags (GPT) ad domains [11]. According to Google, they are are working towards supporting more ad networks. This lack of support for many ad networks along with a limit of 3 ads can significantly impact publishers' revenue. For example, websites that do not use the ad networks supported by Web Light cannot obtain revenue from Web Light users. Moreover, the limit of 3 ads can also constrain the potential revenue of websites showing more than three ads (e.g., by restricting the type of ads and/or their availability across different portions of the webpage that a user may be seeing) from ad networks supported by Web Light. While Web Light pages reduce PLTs, which can *increase* user traffic and thus *raise* advertising revenue, it only does so for supported ad networks on a publishers' pages. It is unclear if the added benefit of increase user traffic out weighs the downsides of limiting ads and only supporting a few ad networks.

To analyze the potential impact of limiting ads and supporting only a subset of the ad networks, we collect Web Light pages and the corresponding original pages, and compare ads shown on them and the ad networks they support.

Methodology. For determining the number of unique ads on the original and transcoded pages, we conduct our experiments in the following way. Using selenium, we open the landing pages of 500 websites from Alexa top 1000 that were being transcoded by Web Light. We also fetch the regular versions of these 500 websites. We repeat this process three times in each case. In each run, we analyze every iframe (if any) on a page recursively and check all embedded links or objects in it (which can be an image source, a link to a JS file or another iframe) and match them with a comprehensive list of ad networks used in Adblock Plus [27]. In case of a match, we count it as one ad. Of course, ads can also be served directly through JS without iframes, however, the latter allows browsers to implement the same origin policy for safety reasons. Thus, our methodology provides a lower bound on ads available on websites.

5.1 Distribution of Ads across Webpages

We find that out of the 500 websites we fetched, only 161 contained ads on their landing pages. Out of these 161 websites, Web Light showed ads on *only* 63 webpages (or 35% of webpages that contained at least one ad) as shown in Figure 7.

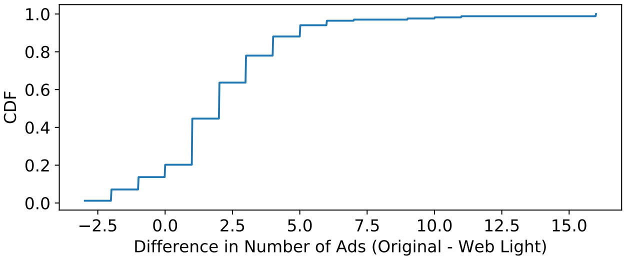
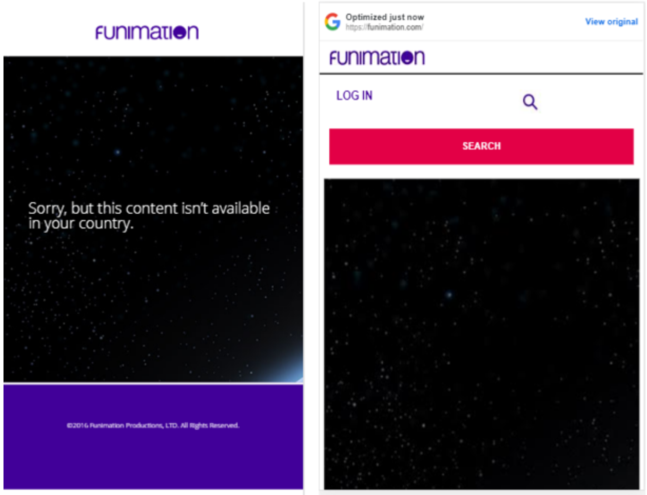


Figure 8: Distribution of the page-wise difference in the number of ads shown on original webpages and their Web Light transcoded versions.



Sorry, but this content isn't available in your country.

©2016 Funimation Productions, LTD. All Rights Reserved.

Figure 9: An example website (https://www.funimation.com/) that does not serve users from Pakistan (left) but is available when accessed through the Web Light service (right).

We find that the maximum number of ads on any regular page is 16 as shown in Figure 7, whereas in case of transcoded pages it is 3 (as also indicated by the Web Light service [11]). Figure 8 shows that for 18% of websites, Web Light showed more ads than the regular counterparts. There are at least two possible reasons for this observation: (1) server-side blocking (also known as *geo-blocking*) or (2) differences in the location of the Web Light proxy and the mobile client (used for fetching the webpage via the direct path).

Geo-blocking and advertising. We test for geo-blocking by first visiting a geo-blocked website via the direct path and later via the Web Light service and then comparing the webpage responses. We find evidence of geo-blocking (also known as server-side censorship) in our experiments, which resulted in differences in ads. For example, https://www.funimation.com/, an entertainment company that specializes in the dubbing and distribution of anime was accessible via Web Light but when accessed via the direct path showed the following message: "*Sorry, but this content isn't available in your country.*" (see Figure 9). Another example is of JCPenny, a department store chain. We found https://www.jcpenney.com/ was accessible via Web Light but the direct path showed the following message, "*We Are Currently Unable to Provide a Shopping Experience for This Country. Please try again later*". The availability of these websites via Web Light may have significant consequences: (a) ads shown to (unintended) users can be wasteful for advertisers and (b) it can increase resource usage and thus costs for provisioning these web services, which did not intend to serve users from certain regions or countries.

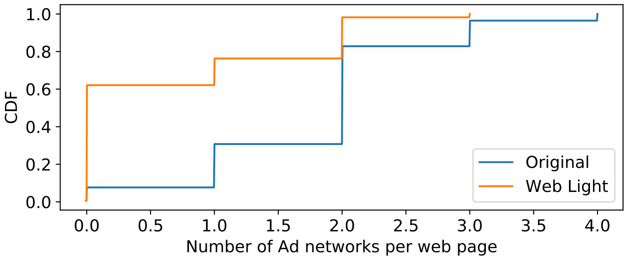


Figure 10: Distribution of the number of unique ad networks on original and their transcoded versions.

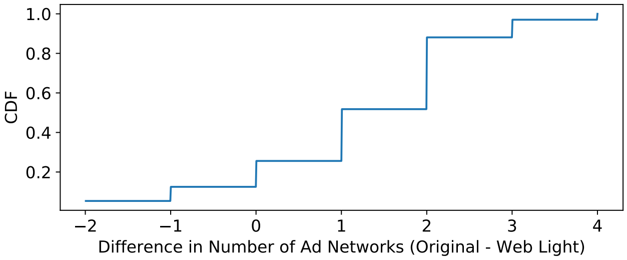


Figure 11: Distribution of the page-wise differences in the number of ad networks in the original and Web Light versions of the pages.

5.2 Distribution of Ad networks

Next, we carry out analysis of *ad networks* used in original pages and their Web Light versions. To conduct this analysis, we simply match the ad domains found on a page with a list of available ad domains used in Adblock Plus. We find that the maximum number of ad networks used in a original webpage in our dataset is 4 whereas 80% of the original webpages served ads via at least two ad networks as shown in Figure 10. In ~50% of webpages, the difference in the number of ad networks in original and Web light pages is at least two implying that either Web Light did not support these ad networks or just removed ads from them because they did not fall in the top three ads (i.e., in the order from top to bottom). For 8% of websites, Web Light webpages served ads from more ad networks than their regular counterparts again suggesting either server-side blocking or location-based variations in ads. Figure 12 shows the popularity of different ad networks across webpages. We find doubleclick.net and googlesyndication.com[5] to be the most popular ad domains.

[5] googlesyndication.com is a domain owned by Google that is used for storing and loading ad content and other resources relating to ads for Google AdSense and DoubleClick from the Google CDN.

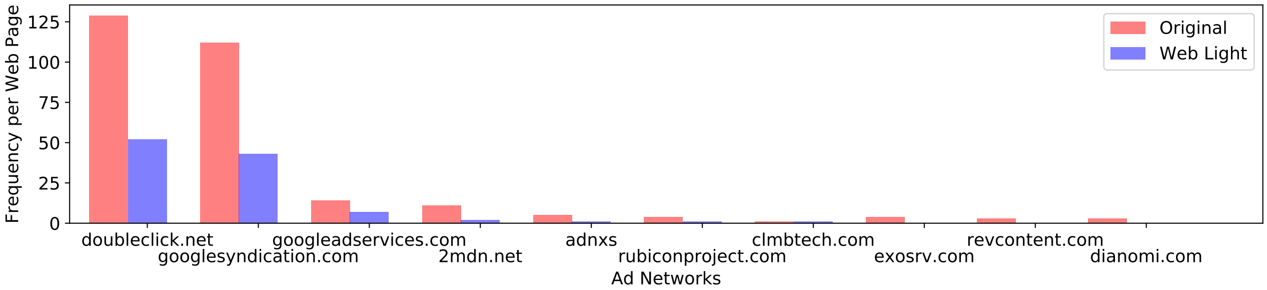


Figure 12: Popularity of different ad networks across the top 500 transcoded pages (from Alexa top 1000) for Original and Web Light transcoded versions. We only show the 10 most frequently occurring pages.

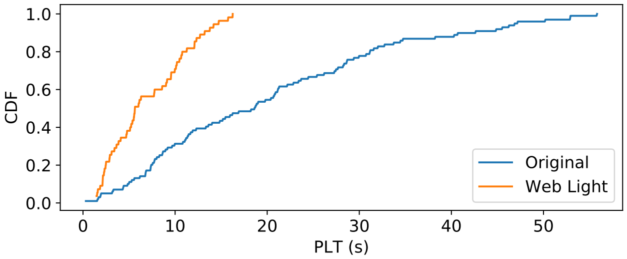


Figure 13: PLTs for Web Light and original pages.

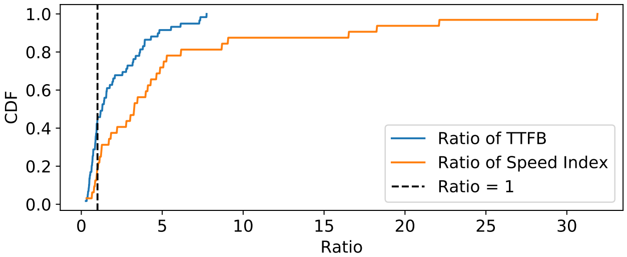


Figure 14: Page-wise ratio of TTFB and Speed Index of the original webpage to the Web Light version.

Which ad networks are supported by Web Light? In the top 500 pages we tested, we found 26 unique ad networks. Out of these, Web Light only supported 9. Ads served from the remaining 17 ad networks were proactively removed when they were the top three ads to appear in a webpage.

Discussion. The FAQ on Web Light's main page (https://support.google.com/webmasters/answer/6211428?hl=en) says, "*Our experiments show that transcoded sites get 50% more traffic than non-transcoded sites and we expect that this will help monetize your site.*". Our results suggest that this may not hold across websites. For example, in case of some websites Web Light showed at least 15 ads less than the original page whereas for multiple sites (e.g., usnews.com and webmd.com[6]) Web Light did not serve *any* ads because it did not support the ad network(s) being used by the webpage to deliver ads. While improving PLTs is known to improve user engagement and thus potentially increase ad revenue, lesser number of ads or lack of support for ad networks can decrease ad revenue. The net effect on ad revenue, which may be negative or positive, will vary across websites.

6 PERFORMANCE WITH WEB LIGHT

We measure the page load performance with Web Light on a Nexus 5 device using a 3G cellular connection. We fetch 100 webpages chosen randomly from Alexa top 1000 webpages.

6.1 Overall performance

We find the median PLT of the original and Web Light transcoded pages to be 19 s and 6.5 s, respectively as shown in Figure 13. We also evaluate the performance based on the SI metric, which is the average time at which *visible* parts of the page are displayed. Figure 14 shows the *page-wise* ratio of the SI of the original page to the SI of the Web Light page. Observe that Web Light reduces the median and 90th percentile speed index by 3.3× and 16.5×, respectively. Interestingly, Web Light increases the PLT (~5%) and speed index (~15%) for some fraction of websites. It happens for two key reasons: (a) for small webpages, Web Light increases the size of their base HTML resulting in more network bytes (as discussed in Section 4 and (b) Web Light's proxy architecture increases path latency and incurs transcoding overhead for uncached pages.

To quantify path latency and the transcoding time at the proxy, we use TTFB as a proxy measure. We find TTFB (which is a function of the network RTT) is larger with Web Light for around 45% of the webpages and can be upto 7.6× larger than directly fetching the page (see Figure 14). This is due to the proxy architecture of the Web Light service, which increases path delay. For the remaining 55% of the pages, Web Light lowers the TTFB because it also caches webpages based on access patterns.

6.2 Impact of Caching

Next, we analyze the impact of caching on Web Light's performance. Web Light generally transcodes a page when a user requests it. While Web Light caches the main content of a page for up to 24 hours[7], pages can be updated more frequently based on users' access patterns. However, Web Light does not cache ads. We find that caching reduces the median PLT by 3.8× over non-cached (or freshly

[6] The Web Light versions of these websites did not show any ads till at least 11am (PST) October 10, 2019.

[7] Other resources such as CSS, JS and images could be cached longer.

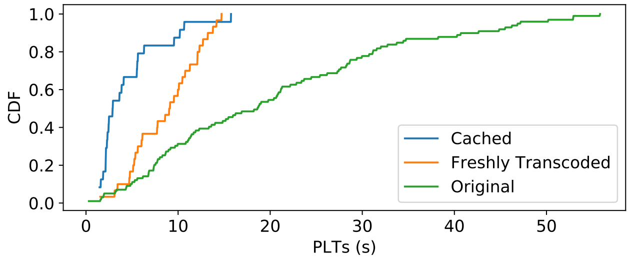


Figure 15: Comparison of Page Load Times between cached, uncached and original webpage.

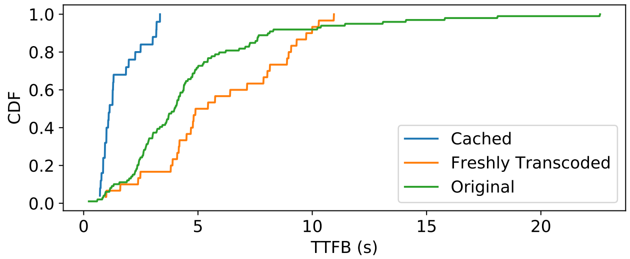


Figure 16: Comparison of Time to First Byte between cached, uncached and original webpage.

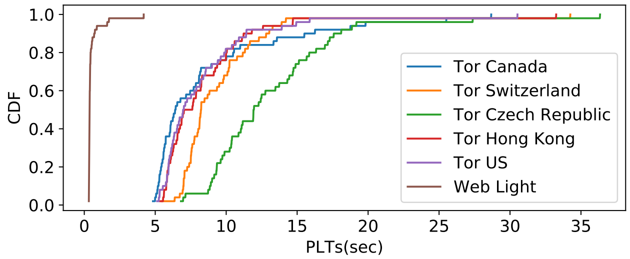


Figure 17: PLTs with Web Light and Tor for a blocked website in Pakistan. For Tor, we use 5 exit relay locations and for each location, we take 50 runs.

transcoded) Web Light pages and by 7.6× over the original pages as shown in Figure 15.

Caching reduces the latency as evidenced by the significant reduction in TTFB as shown in Figure 16, which is likely to benefit small pages. In addition, since the pages are being cached in Google's datacenters, they may reduce the number of hops compared to the direct path, potentially resulting in larger available bandwidth (if the bottleneck is not on the access link).

In summary, Web Light reduces PLTs for a vast majority of websites. Our results show that while the overhead of going through the Web Light proxy (e.g., due to longer path latency and the transcoding overhead) increases the TTFB, the reduction in PLT due to reduction in page size dominates, thereby resulting in an overall reduction in PLTs. Moreover, caching at the proxy provides additional improvements in PLTs (~2× for the median page in our experiments).

7 PRIVACY & INTERNET CENSORSHIP

We now discuss privacy concerns associated with the Web Light service. We then present experimental results that show that Web Light not only provides access to blocked content but also provides *faster* access (e.g., compared to Tor and the Hotspot Shield VPN), which can serve as an incentive for users to use Web Light for censorship circumvention.

7.1 Privacy Concerns

To enable the use of HTTPS pages, Web Light passes the intended URL as a parameter in the proxy's URL. For example, when a user intends to visit https://yahoo.com, it gets sent as https://googleweblight.com/i?u=https://yahoo.com. This has at least two consequences: (a) Web Light is able to show HTTPS pages without breaking them (and without requiring SSL certifcates from the server or requiring the clients to update their root stores) because for the intended server, the proxy service is the client and (b) it implies that Google can potentially see user data in clear text and can easily build user browsing profiles if needed. Thus, such an architecture has significant privacy implications.

7.2 Censorship Circumvention

Internet censorship is prevalent; nearly 70 countries censor Internet content often due to political, social, or economic reasons [26]. As Web Light proxies are hosted in USA, where direct Internet censorship is prohibited by the First Amendment with some exceptions (e.g., child pornography), Web Light also serves as a censorship circumvention tool. We tested 30 blocked websites in Pakistan for censorship. We visited these sites and received a *block* page[8]. However, when using Web Light, all censored pages were accessible. This seems to be an inadvertent use of Web Light but one that can have a substantial effect on users, who have presumably not consented to visiting censored websites and may in fact be unaware of blocked websites in a country. Thus, this can put users in harms way especially in repressive regimes and may also interfere with government regulations and local data privacy laws [21].

Next, we compare the page load performance of censored webpages fetched through Web Light and Tor [23], one of the most widely used tools for anonymous communication and censorship circumvention [23]. While Tor was initially designed as an anonymity tool but in recent years, it has also become a popular circumvention tool. Tor is usually able to circumvent almost all kinds of blocking but fails in regions that block addresses of Tor bridges [26]. Finally, we compare Web Light with the Hotspot Shield VPN, one of the most popular free VPN services with over 650 million users [28].

Web Light provides fast access to censored content. We fetch a censored webpage via Web Light and record the PLT. We then repeat this process with Tor, each time alternating between a Web Light run and a Tor to improve the likelihood that they experience similar network conditions on common path segments. We repeat this process 50 times. In case of Tor, we use different exit relay locations which includes Canada, Switzerland, Czech Republic, Hong Kong, and USA. For each exit relay, we take 50 runs.

[8] This is a page shown to users when they access a censored website to inform them that the desired page cannot be accessed.

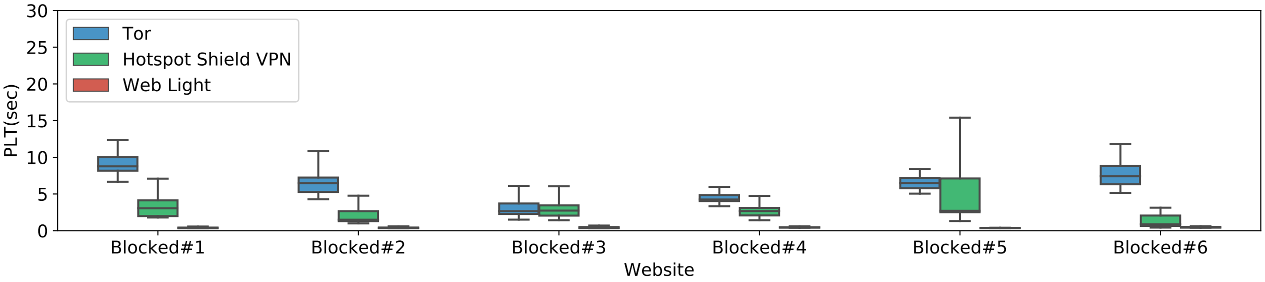


Figure 18: Comparison of Web Light with Tor and Hotspot Shield for six blocked websites in Pakistan. In case of Tor, we pick 5 exit relay locations. The results show that Web Light outperforms both Tor and the Hotspot Shield.

Figure 17 shows the distribution of PLTs with Web Light and Tor for a censored website. Observe that Web Light leads to significantly lower PLTs than Tor across all relay locations. In particular, the median PLT with Web Light was 0.36 s whereas in case of Tor it ranged from 6.4 s, to 11.9 s. We repeat this experiment across six randomly chosen blocked websites from a list of 30 websites. Figure 18 shows the PLT with Web Light and Tor with different exit relays on a box plot. Observe that across all websites, Web Light results in a significantly lower PLTs (e.g., the median PLT improvement ranged from 6.8×-24× over Tor). This happens for potentially the following reasons: (a) Web Light pages have smaller sizes due to transcoding, (b) Tor may have larger path latency as it uses three relays whereas Web Light only uses a single relay, (c) Web Light proxy is hosted in Google datacenters and may have higher capacity to handle load and the path to Google datacenters may also be better provisioned, and (d) Web Light caches pages so after the first request, subsequent requests will see lower PLTs.

Figure 18 also shows that Web Light consistently results in much lower PLTs compared to the Hotspot Shield VPN across all blocked websites. The median improvement in PLTs ranged from 2.4× to 10.1× across the six websites.

8 RELATED WORK

There is a large body of work on improving mobile Web performance, including design of Web proxies [15], understanding dependencies in the page load process [30, 31], studying the impact of network infrastructure [33], and analysis of mobile devices in developing countries [16].

Perhaps, one of the most closely related services to Web Light is Facebook's Free Basics service [3]. Free Basics is available in over 60 countries across select cellular service providers. The Free Basics platform offers a collection of Web services that serve light pages to users so that they can load faster on low-end mobile devices while saving data. Free Basics services do not support JS, large images, or videos. Unlike Web Light, Free Basics pages are pre-screened for compatibility in an offline manner and then allowed on the platform and are separately designed by publishers than regular versions of their websites.

FlyWheel [15] is an HTTP proxy service that reduces data for mobile users by *compressing* responses in-flight between origin servers and browsers. Their optimizations include image transcoding, GZip compression, and JS minification among others. Unlike Web Light, FlyWheel does not provide support for HTTPS pages. Split browser architectures and offloading based systems like Shandian [31] and Opera Mini [6], provide a useful opportunity to improve Web performance by offloading compute intensive tasks such as scripting and rendering. However, they do not support all features (e.g., Opera Mini does not support touch events).

Accelerated Mobile Pages (AMP) is a web component framework by Google for constructing simplified mobile Web pages [1]. The AMP framework consists of three components: AMP HTML, which is standard HTML markup with web components; AMP JS, which manages resource loading; AMP caches, which serve and validate AMP pages. Unlike Web Light, AMP does not target just slow networks and requires rewriting of webpages by the publishers.

9 DISCUSSION

Publisher consent. By default, Web Light attempts to transcode every webpage unless a publisher explicitly opts out. Thus, it transcodes pages *without* publishers' consent. While it seems clear that had publishers been *required* to explicitly opt in before their pages could be transcoded by Web Light, this would have led to a longer time for adoption and perhaps more cost for advertising the Web Light service itself. However, lack of consent leads to several issues around privacy and ethics [24]. In contrast, Facebook's Free Basics service that also uses a proxy architecture, requires *publishers* to themselves come up with a version of their website that complies with a set of guidelines (e.g., a Free Basics cannot have JS, videos, and large images) before they can be accessed via the Free Basics platform. While this addresses the issue of publisher consent, privacy issues still remain with proxy-based architectures that attempt to support HTTPS pages [29].

Monetizing Web services. Web Light complicates monetization of Web services through advertising. For example, it can potentially improve ad revenue by bringing more user traffic to websites as it significantly reduces PLTs. However, since it limits the maximum number of ads to 3 and does not support several ad networks, it can severely limit ad revenues for some websites. It is unclear if the improvement in PLT due to removing ads is worth the potential loss of revenue by publishers, who are arguably in an ideal position to make a decision about such a tradeoff.

Privacy. Web Light's proxy architecture allows Google to potentially build users' browsing profiles and also see users' data in clear text, thereby raising privacy concerns. Moreover, it can also potentially interfere with data privacy regulations such as Europe's

General Data Privacy Regulation (GDPR) [4] and the California Consumer Privacy Act (CCPA) [2].

Censorship circumvention. Using Web Light, users can circumvent Internet censorship. This includes users who have presumably not consented to visiting censored websites and may in fact be unaware of blocked websites in a country. This potentially raises ethical concerns around *informed user consent*, users' privacy and safety [24, 26]. To mitigate potential harm to users, services should inform users about blocked websites in their region and obtain explicit user consent.

Benefits vs. Harms. Our results show that Web Light can substantially reduce PLTs, which can improve mobile QoE and user engagement for users on slow networks. For many such users, the Internet may not be affordable or only partially affordable. For example, according to ITU the mobile broadband price for a 1 GB per month data plan was 14.8% of the average income in least developing countries (or LDCs) compared to 0.8% in developed countries in 2017 [13]. A service like Web Light can potentially bring new users online and also allow existing Internet users to use Internet services for a longer time. On the other hand, there are legitimate concerns around privacy, censorship circumvention, and ad revenue. Whether the potential benefits outweigh the potential harms or otherwise is for users, publishers, and other stakeholders to decide.

10 CONCLUSION

In this work, we presented the first systematic analysis of Google's Web Light service that shows lighter webpages to users by transcoding them on the fly. We found that Web Light significantly reduces page load times by trimming pages and trading off webpage quality, which maybe an acceptable tradeoff for users on slow networks and/or devices. However, it limits publishers' revenue options, raises privacy concerns, and can be used as a censorship circumvention tool. We hope this study brings to light these tradeoffs and helps users and publishers to make informed choices about Web Light.

ACKNOWLEDGMENTS

We thank the anonymous WWW reviewers for their valuable feedback.

REFERENCES

[1] [n.d.]. AMP. https://amp.dev/.

[2] [n.d.]. California Consumer Privacy Act. https://www.caprivacy.org/.

[3] [n.d.]. Free Basics Platform. https://developers.facebook.com/docs/internet-org.

[4] [n.d.]. General Data Protection Regulation. https://eugdpr.org.

[5] [n.d.]. http archive. https://httparchive.org/.

[6] [n.d.]. Opera mini browser. http://www.opera.com/mobile/.

[7] [n.d.]. Secure, private internet access with Hotspot Shield VPN. https://www.hotspotshield.com/.

[8] [n.d.]. Selenium - Web Browser Automation. https://www.seleniumhq.org/.

[9] [n.d.]. Speed Index. https://developers.google.com/web/tools/lighthouse/audits/speed-index.

[10] [n.d.]. Test a website's performance. https://www.webpagetest.org/.

[11] [n.d.]. Web Light: Faster and lighter mobile pages from search. https://support.google.com/webmasters/answer/6211428?hl=en.

[12] 2018. Mobile subscription worldwide outlook. https://www.ericsson.com/en/mobility-report/reports/november-2018/mobile-subscriptions-worldwide-outlook.

[13] 2019. Measuring digital development: Facts and figures 2019. https://www.itu.int/en/ITU-D/Statistics/Pages/facts/default.aspx.

[14] James Titcomb . 2016. Mobile web usage overtakes desktop for first time. https://www.telegraph.co.uk/technology/2016/11/01/mobile-web-usage-overtakes-desktop-for-first-time/.

[15] Victor Agababov, Michael Buettner, Victor Chudnovsky, Mark Cogan, Ben Greenstein, Shane McDaniel, Michael Piatek, Colin Scott, Matt Welsh, and Bolian Yin. 2015. Flywheel: Google's Data Compression Proxy for the Mobile Web. In *NSDI*.

[16] Sohaib Ahmad, Abdul Lateef Haamid, Zafar Ayyub Qazi, Zhenyu Zhou, Theophilus Benson, and Ihsan Ayyub Qazi. 2016. A View from the Other Side: Understanding Mobile Phone Characteristics in the Developing World. In *ACM IMC*.

[17] Android Developers. 2018. Build for Android (Go edition): optimize your app for global markets (Google I/O '18). https://bit.ly/2UKLQDl.

[18] Annkur P Agarwal. 2015. Google Web Light - Move Fast, Break Things? https://pricebaba.com/blog/google-weblight-move-fast-break-things.

[19] Barry Schwartz. 2015. GoogleWebLight.com: Google's Answer to making your website faster. https://www.seroundtable.com/google-google-web-light-20558.html.

[20] Barry Schwartz. 2018. New Google Search Console adds filter for Web Light search appearance. https://searchengineland.com/new-google-search-console-adds-filter-for-web-light-search-appearance-300206.

[21] S. Burnett and N. Feamster. 2015. Encore: Lightweight Measurement of Web Censorship with Cross-Origin Requests. In *ACM SIGCOMM*.

[22] Chromium and WebDriver teams. [n.d.]. ChromeDriver – WebDriver for Chrome. https://chromedriver.chromium.org.

[23] Roger Dingledine, Nick Mathewson, and Paul Syverson. 2004. Tor: The Second-generation Onion Router. In *USENIX Security Symposium*.

[24] D. Dittrich and E. Kenneally. 2012. *The Menlo Report: Ethical Principles Guiding Information and Communication Technology Research*. Technical Report. U.S. Department of Homeland Security.

[25] Erin Acheson. 2018. Web Light Appearance Update for Google Search Console. https://www.demandsphere.com/blog/web-light-appearance-update/.

[26] Aqib Nisar, Aqsa Kashaf, Ihsan Ayyub Qazi, and Zartash Afzal Uzmi. 2018. Incentivizing Censorship Measurements via Circumvention. In *ACM SIGCOMM*.

[27] Peter Lowe. 2020. Blocklist for use with Adblock Plus. http://pgl.yoyo.org/as/serverlist.php?hostformat=adblockplus.

[28] Rob Mardisalu . 2020. Hotspot Shield Review. https://thebestvpn.com/reviews/hotspot-shield/.

[29] Rijurekha Sen, Sohaib Ahmad, Amreesh Phokeer, Zaid Ahmed Farooq, Ihsan Ayyub Qazi, David Choffnes, and Krishna P. Gummadi. 2017. Inside the Walled Garden: Deconstructing Facebook's Free Basics Program. *SIGCOMM Comput. Commun. Rev.* 47, 5 (Oct. 2017), 12–24.

[30] Xiao Sophia Wang, Aruna Balasubramanian, Arvind Krishnamurthy, and David Wetherall. 2013. Demystify Page Load Performance with WProf. In *NSDI*.

[31] Xiao Sophia Wang, Arvind Krishnamurthy, and David Wetherall. 2016. Speeding up Web Page Loads with Shandian. In *NSDI*.

[32] Worldometers. [n.d.]. Countries in the world by population. https://www.worldometers.info/world-population/population-by-country/.

[33] Yasir Zaki, Jay Chen, Thomas Pötsch, Talal Ahmad, and Lakshminarayanan Subramanian. 2014. Dissecting Web Latency in Ghana. In *ACM IMC*.