Received October 21, 2019, accepted November 6, 2019, date of publication December 13, 2019,
date of current version December 23, 2019.

Digital Object Identifier 10.1109/ACCESS.2019.2959731

An Ad-Driven Measurement Technique for Monitoring the Browser Marketplace

PATRICIA CALLEJO[1,2], RUBÉN CUEVAS[2,3], AND ÁNGEL CUEVAS[2,3]
[1]IMDEA Networks Institute, 28918 Leganés, Spain
[2]Telematic Engineering Department, Universidad Carlos III of Madrid, 28911 Leganés, Spain
[3]UC3M-Santander Big Data Institute, 28903 Getafe, Spain

Corresponding author: Patricia Callejo (patricia.callejo@imdea.org)

This work was supported in part by Ministerio de Economía y Empresa, Spain, under the Grant RyC-2015-17732, and in part by the European H2020 projects SMOOTH (786741) and PIMCITY (871370).

ABSTRACT In this paper we present a novel active measurement methodology for monitoring the browser market landscape. It leverages the display ads delivered through online advertising campaigns to collect the browser brand and version of the device receiving the ad. While providing a similar accuracy to traditional techniques based on passive measurements, our methodology offers some advantages: (*i*) a lower entry barrier for researchers and practitioners interested in measuring the browser marketplace; (*ii*) it allows targeted measurements, which can be useful to fix biases in the data sample or to analyze specific aspects of the browser market. We analyze the performance, accuracy, and capabilities of our methodology through real experiments that overall produced more than 6M measurements.

INDEX TERMS Active measurements, ads, browser market place.

I. INTRODUCTION

In the current Internet, desktop computers interact with a large number of services, including the most popular ones (Online Social Networks, Video Portals, Streaming services, etc.), through browsers. Although this affirmation is not valid for mobile devices where most popular services run through proprietary applications; we cannot deny the importance of browsers in the mobile ecosystem as well. Arguably, we can assert that browsers are the most important online tool on the Internet, used every day by billions of users.

Having the control of a widely used browser helps to bring a technology company into a privileged position. For instance, a company with a dominant position in the browser market can (among other things):

- Influence the adoption of different web technologies (e.g, flash vs. HTML5).
- Have access to the browsing history, and thus accurate information, about the interests of hundreds of millions of users. Such information is precious for digital marketing (a business generating a revenue of $107.5B in 2018 just in US [1]).

The most important technology companies (Apple, Google, and Microsoft) are aware of the importance of having a strong position in the browser market so that they dedicate a large amount of resources to develop their browsers. Measuring the browser market share is, for obvious reason, relevant for these browser development companies. However, it is also important for other businesses such as:[1]

- Software development companies including online gaming companies, e-commerce sites, plug-in development companies, benefit from knowing the browser marketplace so that they are aware of the most popular browser brand and version per region and demographic end-users profile (i.e., age and gender) and thus the most critical for their own business.
- The security bugs of a browser's version are typically reported and fixed in the next released version [2], [3], and thus vulnerable security browsers are typically associated with old versions. Online security companies commercializing products such as firewalls, antivirus, etc., know the security bugs and vulnerabilities of different browsers' brands and versions. However, they do not know how widespread are such vulnerable browsers or identify where they are located (e.g., country, IP prefix,

The associate editor coordinating the review of this manuscript and approving it for publication was Feng Xia.

[1] Note that this is a non-exhaustive list of businesses benefiting from using a solution to measure the browser landscape market share.

Internet Provider). Therefore, a tool to quantify the presence of vulnerable browsers, and identify where they are, is of great value for these companies to control the damage these versions may cause.

- Digital marketing companies run display campaigns whose ads are shown in browsers and mobile apps. Knowing the browser market share per geographical region and end-users demographic profile (age and gender) would help them: i) from a marketing perspective define better targeting strategies based on the profile of users associated to different brands and versions; ii) from a technical perspective, they can make sure their scripts (creativities) run (render) correctly in those most popular browsers brands and versions.

In addition to these business reasons, there are other important arguments to develop these techniques, for instance:

- Having an independent and accurate estimation of the browser market share would allow regulators in different parts of the world to assess the presence of monopoly situations in such critical aspects as the browser landscape. Indeed, the European Union takes monopoly situations very seriously and have already sanctioned Google for abusing its dominant position [4].
- The utilization of different browsers has different associated implications. Browsers owned by companies related to the business of digital marketing (e.g., Google or Microsoft) may collect end-users information for their digital marketing products. Other browsers like Firefox are supported by non-profit foundations, i.e., Mozilla, with no commercial interest. Therefore, the use of these types of browsers provides, in principle, higher privacy guarantees [5].

A monitoring solution should account with three principal characteristics to offer the functionality described above:

1) Scalability: It should be able to retrieve a large scale sample with at least millions of data points in order to provide statistically representative results.
2) Accessibility: Any entity including small private companies, researchers, regulators should be able to use and obtain valid results from it.
3) Geographical and Demographic Targeting Capability: The solution should offer the possibility to run targeted measurements on specific geographical locations (e.g., a country or a region) and for specific demographic groups of users based on their age and their gender. By doing so, for instance, a company willing to lunch a new online software targeting a specific demographic group (e.g., male between 20 and 30) in a given geographical location (e.g., France) can know the browser market share of its target in advance and chose the best development strategy to follow.

Existing solutions to monitoring the browser market landscape rely on passive measurement techniques [6], [7]. The monitoring company installs a tracking code in a pool of N websites to collect the browser id associated with each visit to these websites. N ranges between thousands and millions of websites. While this methodology provides the scalability property and thus it is undoubtedly valuable to provide a solid knowledge about the browser marketplace, it presents important limitations in the accessibility and targeting capabilities since (*i*) it is accessible to just few companies with capacity to monitor thousands of websites and (*ii*) its passive nature prevents targeted monitoring campaigns for specific geographical areas or browser ids/versions. In addition to commercial solutions, there are academic works that analyze the browsers' marketplace with a focus on security [8], [9]. However these works do not develop any specific technique to collect browser information, instead they use logs from Google, which is obviously proprietary and not accessible.

In this paper, we propose the use of AdTag [10], an active measurement platform leveraging display ads as measurement vantage points, to monitor the browser market landscape. This approach overcomes the described limitations of traditional passive methodologies and offers the three main functional requirements mentioned above: scalability, accessibility and targeting capabilities. AdTag inserts a lightweight JavaScript code within display advertisements. When an impression of these instrumented ads is displayed on a website, the embedded JavaScript code collects the User-Agent and the IP address[2] of the device. The User-Agent reveals the browser brand and its version, whereas the IP address allows us to map the device to a geographical region.

There is a large number of advertising providers that can serve as an appropriate infrastructure to execute AdTag. Moreover, the cost associated with it is low. Note that the price of a thousand impressions (a.k.a. CPM) can be as low as $0,01 in some providers. For instance, the provider used in our experiments has a CPM starting at $0,10. This shows that our technique is accessible and affordable for any institution (small private companies, researchers, regulators, etc.) interested in monitoring the browsers marketplace. Moreover, it offers the required scalability, allowing to obtain millions of measurements per day with a low investment of tens to hundreds of dollars.

In addition, our proposal can leverage the targeting capacity of the online advertising ecosystem to set up targeted measurement campaigns based on geographical location and demographic properties (age and gender). Note that other targeting parameters are also available, e.g., Operating System, device type (mobile vs. desktop vs. tablet), etc.

Finally, to prove the efficiency of our methodology, we have run general purpose as well as targeted experiments:

- Our general purpose experiments were configured without targeting parameters. We collected more than 6M measurements. We compare the browsers' market share obtained from our measurements and the one reported by well-known companies using traditional passive

[2] Note that we use the IP address to extract metadata information. Afterward, we anonymize the IP using hashing techniques.

techniques. The results indicate that the discrepancy between our results and those reported by companies using traditional techniques is in the same range as the discrepancy between the results of these companies. Therefore, we conclude that our methodology provides reasonably accurate results.
- To exemplify the targeting capacity of our methodology, we run three targeted campaigns: First, a geographically targeted campaign for Albania, a location with no representation in our general dataset. Second, a targeted campaign focused on old versions of operating systems, which are likely to run outdated versions of browsers with security vulnerabilities. We are able to discover the presence of more than 30 thousand devices, out of 345k, using outdated browsers with security vulnerabilities within 3 days of duration of our campaign. Third, a demographically targeted campaign for people aged between 18 and 25 years old in Italy on mobile devices. This fact shows that our proposed methodology presents functionalities not available in the traditional passive technique.

The remainder of the paper is organized as follows: Section 2 describes the traditional monitoring technique and its drawbacks. Section 3 presents our active measurement approach and discusses its advantages and limitations. Section 4 shows the empirical results obtained from applying our methodology in general purpose and targeting use cases. Finally, Section 5 concludes the paper.

II. TRADITIONAL METHODOLOGY FOR MONITORING THE BROWSER MARKETPLACE

A. OVERVIEW

The traditional methodology for monitoring the Browser Market landscape consists of installing a tracking code in a large number of websites. This code collects the User-Agent associated with each visit. The User-Agent serves as a browser identifier which reveals the browser's brand and version. Moreover, the tracking code can collect some other information such as the IP address of the device visiting the website. The IP address can be processed to retrieve the geolocation of the visit.

Several companies implement this traditional methodology: StatCounter, W3Counter, Net Applications, Wikimedia, etc. Each of them has access to a different set of websites and thus report their results based on an independent set of visits. These companies monitor between thousands (W3Counter [6] or Net Market Share [11]) and millions (StatCounter [7]) of websites. These companies provide datasets including up to 15B visits per month as in the case of StatCounter.

B. LIMITATIONS

Next we discuss the main limitations of the described traditional methodology:

- High Entry Barrier: The described traditional methodology presents very high entry barriers, limiting its use to a handful of companies able to install their monitoring code in (at least) thousands of websites. For instance, the research community is (in general) excluded from the use of this methodology since it is very unlikely that a research team can have access to a such large pool of websites.
- Bias in data samples: Traditional techniques register the visits to their monitoring websites as data samples. This means that a user visiting 50 times with the same browser a web page generates 50 data inputs. This is potentially a source of bias, since heavy visitors to the monitoring websites would have a higher weight in the final market distribution across brands and versions. Furthermore, the reach of this methodology is limited to the users that visits those websites, which can not be the real distribution of the browsers' market share.
- Other biases: Despite each of the companies using the traditional methodology accounts with a large data sample, we find discrepancies among their reported browsers' market share. The reason is that the collected dataset may be affected by different biases: geographical biases (having a higher representation from users located in certain countries), OS biases (having a higher representation from users of a specific Operating System), Type of device bias (having a higher representation of mobile or desktop devices), etc. These biases are associated with the pool of websites used by each company. For instance, if the pool of websites is predominantly in a specific language (e.g., English), there will be a geographical bias towards countries speaking such language.
- Does not allow targeted measurements: The vision of the browser marketplace depends on the users that connect to the monitoring websites. This fact is out of the control of the company. Hence, even if a bias is identified in the sampled data (e.g., a geographical bias), the company has many difficulties for fixing it because they cannot modify the demographic or geographic properties of the user base connecting to the monitoring websites. Instead, if the company had some capacity to define the targeting population for their measurements, it could fix identified biases in a simpler manner.

III. ACTIVE MEASUREMENT BASED METHODOLOGY FOR MONITORING THE BROWSER MARKET LANDSCAPE

A. OVERVIEW

Our goal is to define a methodology that overcomes the limitations of the traditional techniques for monitoring the browser marketplace discussed in the previous Section. In particular, this methodology should meet the following requirements: (*i*) low entry barrier so that any person/company/research team interested in monitoring the browser market can do it at a reasonable cost; (*ii*) it should allow targeted measurements. This will serve to fix identified biases in the data sample, but also to conduct specific analysis of the browser market landscape such as analyze the market share for a particular demographic group (based on age and/or

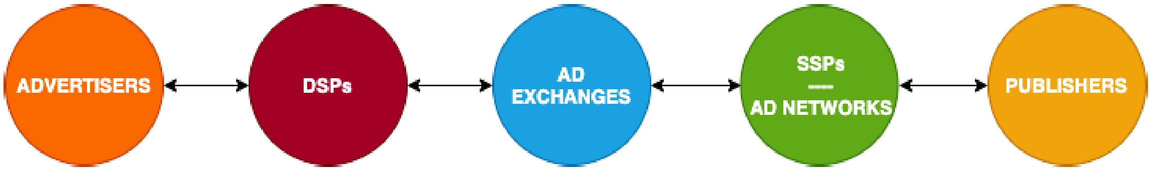

FIGURE 1. Overview of the programmatic adverting ecosystem.

sex), analyze the presence of insecure browser versions and identify its associated IPs, etc; (*iii*) it should guarantee that each browser instance represents a single data sample in the collected dataset.

To achieve these goals, we propose the utilization of AdTag [10], and active measurements platform, contrary to the passive measurements used so far. In particular, our approach relies on the online advertising ecosystem. Most websites have embedded ads, we propose to use these ads as vantage points to collect the User-Agent and IP address of the device connecting to webpages where an ad, we have under control, is shown. It is estimated that around a trillion ads are delivered every day. This number provides a solid basis to meet the required scalability to monitor the web browser marketplace. Moreover, the online advertising ecosystem offers the needed functionality to achieve the goals described above. First, it allows running targeted advertising campaigns based on different parameters including: demographic characteristics (age and sex of the user), geographic location (country, region, and even cities), type of device (desktop vs. mobile), Operating System, etc. Second, any person or company can use one of the hundreds of available online advertising vendors to configure their own campaigns using the monitoring methodology described in this Section. Third, online advertising campaigns offer a configuration parameter referred to as Frequency Cap, which determines the maximum number of times an ad is shown to a specific user, i.e., browser in our case. By setting up the Frequency Cap equal to 1, we obtain solid guarantees that a specific browser instance will contribute a single data sample to our dataset. Finally, the experiments have a low cost. As a reference, the cost of 1M measurements ranges between \$10 and \$100 approximately, depending on the vendor. Therefore, the entry barrier of our proposed methodology is significantly lower than the one imposed by the traditional methodology.

In the remainder of the Section, we first present a brief overview of the functionality of the Online Advertising Ecosystem, which will allow understanding better the details of our methodology, presented afterward. Subsequently, we discuss the limitations of our proposed methodology and ways of addressing them. Finally, we discuss the ethical aspects to be considered when using the proposed methodology.

B. ONLINE ADVERTISING ECOSYSTEM

The online advertising ecosystem has evolved from the so-called private markets, where an advertiser signs private contracts with the websites to show its ads to a more complex system referred to as programmatic advertising. In programmatic advertising, websites lease their ad spaces to an ad network. This ad network is responsible for finding an advertisement to be placed in the leased ad space. Upon the reception of a new visitor, the webpage (called publisher) sends an ad request along with complementary information such as the IP address, the UA, and cookies to the ad network. Using this complementary information, ad networks are able to enrich the ad request, including information about the location of the user, the device type, OS, browser (based on the IP and the UA), and the user's interests and demographic profile (based on cookies). This enriched ad request is passed to an Ad Exchange. The Ad Exchange is an entity that puts in contact advertisers willing to deliver ads with ad networks in programmatic advertising. The Ad Exchange maps the received enriched ad request into a bid request, which is sent to the advertisers. Based on the information included in the bid request, interested advertisers generate a bid for a given price to participate in an auction process handled by the Ad Exchange. Upon the reception of all the bids, the Ad Exchange chooses a winning bid. The advertiser that issued the winner bid delivers its ad to the user. The described bidding process is performed automatically. In particular, to automate this process, advertisers use Demand Side Platforms (DSPs), which connect advertisers to Ad Exchanges and issue the bids on behalf of advertisers.

The revenue generated by an advertiser's won bid is shared between the DSP, the Ad Exchange, the ad network, and the website. Figure 1 shows a visual representation of the Online Advertising Ecosystem functionality. Note that we have described a simplified version of the actual programmatic ecosystem to allow readers to understand the rest of the paper. However, the programmatic ecosystem is more complex and involves more players than those described in this document. Interested readers can find further information in [12].

C. DETAILS OF THE METHODOLOGY

Online advertising offers different forms of ads: video ads, display ads, search ads, etc. AdTag leverages display ads. These are the typical banner ads that appear on most websites. Display ads are currently developed in HTML5. Then, they can include JavaScript code. We take this opportunity to insert a custom JavaScript code to collect the User-Agent information.

We create our own HTML5 ad, which includes a JavaScript code for collecting the User-Agent information. We set up an advertising campaign with our instrumented ad in a DSP.

Once this campaign is started each time an impression of our ad is delivered, the JavaScript code retrieves the User-Agent of the browser receiving the ad. Then, the JavaScript code establishes a TCP connection with a central server where we store the collected information. The server obtains the IP address of the device receiving the ad from this TCP connection.[3]

Therefore, each time our ad is displayed we collect a tuple including the following information: <*timestamp, IP address, User-Agent*>.

Each of these tuples is processed. We use the GeoLite MaxMind database[4] to map the IP address to a geographical location (country and region) and two Python libraries, `user_agents`[5] and `httpagentparser`[6] to map the User-Agent to its browser brand and version as well as to obtain the OS and OS's version. After this, the IP address is anonymized using hashing techniques. Therefore the final tuple stored in a central database is: <timestamp, hashed IP address, country, region, browser's brand, browser's version, OS, Os's version>.

Finally, this methodology allows performing active targeted measurements. As we have described earlier, an advertiser can configure display ad campaigns targeting specific audiences, which are defined by a combination of geographical location, demographic characteristics, users' interests, device type, operating system, etc. These options are available in most DSPs.

D. PERFORMANCE EVALUATION

We have run our server code on a standalone machine (24 2.4GHz cores, 64GB RAM). Under this setting, it is able to handle over 100k simultaneous connections. Note that in case more resources are needed, multiple servers can be installed using load-balancing techniques to distribute the load among them. Therefore, the proposed methodology offers the necessary scalability to collect (at least) hundreds of millions of measurements every day.

Moreover, our methodology is meant to run in the wild through real ad campaigns. Hence, it may be affected by different type of errors, which may prevent collecting the information from some ad impressions: browser extensions preventing the deployment of ads or JavaScript code (e.g., ad blockers [13] or no-script [14]), network problems preventing the establishment of the connection, problems in the execution of the JavaScript code of our ad, etc. We have observed that on average our methodology was not able to collect information for 15% ad impressions. This rate was computed as the ratio between the number of ad impressions recorded with our methodology and the total number of ad impressions reported by the DSP used to run the ad campaigns. A careful analysis of these losses indicates that most of them are due to the fact that the server used in this paper was running on an academic network that offers good performance, but it is not designed to support large-scale experiments receiving a large number of connections. Indeed, we have run our methodology for a different research project within the infrastructure of an ad-tech provider network (this one designed to handle a large number of connections) experiencing a much lower fraction of losses below 5%.

E. LIMITATIONS

In this section we discuss the main limitations of our methodology with respect to the traditional passive measurement techniques.

- Scalability: Some well-established companies using the traditional methodology can reach in the order of millions to hundreds of millions measurements per day. However, just a handful of companies have the coverage and infrastructure to reach such scale. Our methodology has the theoretical capacity to achieve such magnitude, but it would require a recurrent high investment. If we assume a CPM of \$0.10, obtaining 1M (100M) daily measurements would cost \$100 (\$10000). Therefore, reaching equivalent scalability as the one offered, for instance, by StatCounts seems unfeasible due to the high economic cost. However, reaching scalability in the order of a few millions of measurements per month is affordable for interested companies or research teams. As we will show in section IV, a few million measurements suffice to obtain results similar to those presented by companies using the traditional methodology accounting with billions of measurements every month.
- Data sample biases: As in the case of traditional measurements, our methodology is subject to suffer from biases in the obtained data sample (e.g., underrepresented geographical areas or demographic groups). However, once the bias is identified, our methodology allows taking correction measures by defining complementary ad campaigns that target the underrepresented audiences. This is a clear advantage over the traditional technique, which cannot take straightforward countermeasures to existing biases in its data sample.
- Device Resource Consumption: Our methodology requires the device receiving the ads to devote some computation resources to execute the JavaScript code and some bandwidth to send the collected information to the central server. Contrary, the traditional methodology does not require to use any resource from the device, since it uses passive measurements. We have carefully evaluated the resource consumption in lab experiments. The computation resources used by our JavaScript code are negligible (executing a call to the browser API to retrieve the User-Agent and establishing

[3] Note that an alternative and more lightweight manner of doing this is sending an HTTP GET message to the server. However, certain ad-tech providers block GET requests if they come from a third party.

[4] http://www.maxmind.com

[5] https://pypi.org/project/user-agents/

[6] https://pypi.org/project/httpagentparser/

a TCP connection). Moreover, our JavaScript code sends a message of 600 Bytes to the central server. Hence, the consumption of end-users' data is also minimal.

F. ETHICAL CONSIDERATIONS

The leading author has conducted the experiments described in this paper. The Ethical Committee of the leading author's institution reviewed the research as well as the experiments and provided the approval for their execution since they do not incur in any unethical action. The two main ethical concerns associated with the experiments relate to the retrieval of the IP address, which is considered PII by the EU legislation [15] and the consumption of data in the communication of the ad tag with the server. As described above, the IP address is anonymized after extracting its meta-information while the data consumption is very limited and negligible in comparison with the overall data consumption associated with loading a webpage.

G. IMPLICATIONS FOR BUSINESSES

As discussed in the introduction, having an accurate solution to estimate the browsers' market share is important for several companies, including different types of software development companies, online security firms, companies operating in the digital marketing ecosystem, etc. One of the main problems of traditional solutions is that, as mentioned above, they require a monitoring infrastructure only available to a few companies. Indeed, most companies in the mentioned businesses (software development, online security, and digital marketing) do not have such infrastructure. However, any of them can use our proposed solution, since it does not require to have any pre-existing large-scale infrastructure and can be deployed on-demand through any of the hundreds of available providers. Therefore, our solution allows, for the first time, the democratization of the monitoring of the browser marketplace to any company regardless of its size.

Second, our technique provides targeting capabilities, not offered by traditional solutions. This offer companies the possibility to perform specific studies based on their own needs. For instance, as we will show in section IV, a security company can use these targeting capabilities to identify installations of vulnerable browsers and take appropriate protection measures. Also, we will show how these targeting capabilities would allow a company to perform an accurate study of the browser market share in underrepresented geographical areas (e.g., a country). This could be, for instance, useful for a software company developing a web application in one of such geographical areas in order to identify the most popular browser versions and make sure the web application works properly for them. Finally, the demographic targeting capabilities can be again leveraged by businesses to understand the browser market share across the targeted demographic group. For instance, a company developing an online game for males between 20 and 30 years in a specific geographical area (e.g., France), can use our solution to characterize the browsing marketplace in that region and optimize the performance of the game for those browsers most commonly used by the targeted demographic group.

In summary, companies interested in understanding and characterizing the browser marketplace can benefit from our solution for two main reasons: 1) its accessibility for companies of any type and size and 2) its targeting capabilities.

IV. EXPERIMENTS AND RESULTS

In this section, we evaluate the ability of the proposed methodology to monitor the web browser marketplace. To this end, we first present the results of running a large-scale general purpose non-targeted campaign. This campaign serves to assess the accuracy of the estimation of the browser market share reported by our methodology in comparison with the reports of companies using the traditional methodology. Afterward, we present the results of two targeted experiments, whose goal is to showcase the targeting capacity of our methodology. In particular, we run a geographically targeted experiment focused on Albania (a country underrepresented in the general purpose dataset). Secondly, we run a targeted experiment focused on identifying the presence of outdated versions of browsers presenting security vulnerabilities. To this end, we configure targeted campaigns to old versions of OSes, which are likely running obsolete versions of browsers. Finally, we run a demographic campaign to identify the most popular browsers used among a specific audience. To achieve this goal we configure a targeted campaign for people between 18 and 25 years old, in Italy, using mobile devices.

A. MEASUREMENT PLATFORM AND EXPERIMENT SETUP

We configure our ad campaigns through a DSP, which allows us to set up the following targeting parameters: geographical location, device type, OS brand, OS version, specific User-Agent, demographical information, etc. From these, in this paper, we use only three targeting parameters, the geographical location, the demographical information, and the OS brand and version. Next, we describe the specific set up of each of the four run experiments:

1) LARGE-SCALE NON-TARGETED CAMPAIGN

We configure a campaign in which we do not select any targeting parameter. This campaign is run through 9 well-known vendors including Google, AOL, Pubmatic, etc. We have run this experiment twice in May 2017 and Oct 2017, generating more than 3M measurements in each of the experiment. We present the combination of both datasets for the results of this paper.

2) GEOGRAPHIC-TARGETED CAMPAIGN

We use the large-scale dataset obtained in the previous experiment to identify underrepresented countries (i.e., countries with a very low number of samples) and chose one of them to run a targeted ad-campaign to show how our methodology (contrary to the traditional one) can take actions to correct biases. To this end, we configure a targeted campaign to

TABLE 1. Browsers' market share obtained from our general purpose large-scale dataset. Results show the absolute number of samples and its equivalent percentage per OS.

Browser	total
Chrome	2435760 (40.85%)
Chrome Mobile	1262945 (21.18%)
Safari Mobile	898924 (15.08%)
Firefox	428896 (7.19%)
IE	277716 (4.65%)
Safari	181277 (3.04%)
Edge	133241 (2.23%)
total	5618759 (94.22%)

deliver ads to Albania. We obtain a total of 3k measurements in a couple of days (note that in the large-scale dataset we only have 14 entries from Albania, in one week).

3) OS-TARGETED CAMPAIGN

Our goal is identifying outdated browsers with severe security vulnerabilities, which represent serious security threats. Note that our methodology would allow to identify the IP addresses associated to those browsers.[7] As a result, the Security responsible of the institution or provider hosting such an IP can be warned. To identify these type of browsers we configure ad campaigns targeting old versions of OSes, in particular we target Windows XP, Mac OS X v10.0 (known as Cheetah) and Linux v686. Instances of outdated browsers are likely to run in old version of OSes. As a result of this experiment we obtained a total of 345k measurements.

4) DEMOGRAPHIC-TARGETED CAMPAIGN

To showcase the demographic targeting capabilities of the proposed solution, we have configured a campaign targeting the following demographic group: young people (ages between 18-25) in Italy using mobile devices. We obtain a total of 13k data samples that provides an estimation of the mobile devices and browser market share across the targeted demographic group.

The overall cost of all these experiments was around \$720 at an average CPM of \$0.5. These numbers offer a cost reference that confirms that our methodology presents a low entry barrier in comparison with the traditional methodology that requires direct access to a large number of websites.

B. RESULTS

1) ACCURACY OF ESTIMATION OF BROWSER MARKET SHARE

Using our large-scale dataset, we have computed the market share of different mobile and desktop browsers. Results are presented in Table 1. Moreover, Table 2 shows the list of countries where each of the three most common browser brands (Chrome, Safari, and Firefox) shows the highest presence. We merged the desktop and mobile platforms for this table.

TABLE 2. List of countries where each of the major browser brands (Chrome, Safari and Firefox) have highest presence.

Browser	Country	total (%)
Chrome	United States	865299 (23.39)
	Brazil	289045 (7.82)
	United Kingdom	276000 (7.46)
	Turkey	221751 (5.99)
	Canada	191284 (5.17)
	Mexico	159655 (4.31)
	France	136857 (3.7)
	Argentina	118946 (3.2)
Safari	United States	432906 (40.07)
	United Kingdom	191143 (17.69)
	Canada	85938 (7.95)
	Australia	48396 (4.48)
	France	43137 (3.99)
	Ireland	32306 (2.99)
	Netherlands	27645 (2.56)
	Germany	23351 (2.08)
Firefox	United States	117141 (26.62)
	United Kingdom	39722 (8.91)
	Germany	38708 (8.68)
	France	31652 (7.10)
	Poland	24625 (5.52)
	Canada	17794 (3.99)
	Brazil	16844 (3.78)
	Spain	13130 (2.95)

TABLE 3. Market share reported by StatCounter, W3Counter, and NetMarketShare compared with our methodology, AdTag.

Browser	StatCounter	W3Counter	NetMarketShare	AdTag
Chrome	62.7%	57.4%	63.88%	64.88%
Safari	15.89%	13.5%	17.64%	18.15%
Firefox	5.07%	6.8%	4.76%	7.22%
IE & Edge	4.68%	6.8%	5.72%	7.05%
Opera	2.55%	2.4%	0.89%	0.92%

Finally, Table 3 compares the market share reported by our methodology and other companies using the traditional methodology. If we take StatCounter as a reference for comparison (since it is the company accounting with a larger sample of data), we observe that our results present an average difference of 4 percentage points across the different browser brands. This difference is equivalent to that shown by other companies using the traditional methodology. In particular, StatCounter and NetMarketShare show an average difference of 3 percentage points.

The differences of reported results across different systems are caused by the distinct biases present in each dataset. In the lack of ground truth, it is not possible to conclude which report presents the closest results to such ground truth. However, all reports, including ours, show a high coherence, indicating that all of them are reasonably accurate.

Therefore, despite the fact that our methodology cannot reach, in practice, the volume of samples that some companies achieve using the traditional methodology, we can conclude that the results accuracy is expected to be similar as the one achieved by the traditional methodology.

2) GEOGRAPHIC TARGETING

Our general purpose measurement campaign shows several underrepresented countries. One of them is Albania that

[7] For ethical reasons we only stored an anonymized version of the IP address.

contributes just 14 data samples out of the 3M. In the traditional methodology, such events cannot be easily addressed since the composition of the data sample is not under the control of the company issuing the measurements. To address this geographical bias under the traditional methodology, a company may try to add to its websites' pool the most popular pages in the underrepresented countries. This action may take time (it requires reaching the administrator of the website, establishing a negotiation, a (economic) compensation, etc.) and the success is not guaranteed. Instead, our methodology can straightforwardly address this type of biases since we can define targeted campaigns focused on specific geographical locations. The results of our geographical-targeted campaign proves it. We have run this campaign during 3 days, obtaining 3k data samples from Albania, addressing the under-representation problem of this country.

3) SPECIFIC BROWSER TARGETING

The described experiment produced a total of 345k data samples, distributed across pairs of browsers/versions we will analyze further. By the time we run this experiment, the stable versions of Chrome, Safari, and Firefox were 63, 59, and 11, respectively. For security reasons, all browsers recommend to update the engine to the latest version, and they have by default the option to update the version automatically. We consider that any browser with a version (at least) 4 years older than the mentioned version are likely linked to security vulnerabilities and thus, represent a security threat. Table 4 show the number of impressions served to browsers versions lower than the ones mentioned above.

While a general purpose measurement as the one in our large-scale experiment is able to identify some of these insecure browsers (25k out of the 5.2M of data samples from Chrome, Firefox. and Safari), a targeted study helps to unveiled a much larger number of them, as we demonstrated in this experiment. The traditional methodology does not have this capacity since its passive nature limits its ability to select a measurement target.

We are aware that some browser versions less than 4 years old are also vulnerable. However, the goal of our experiment is not to identify all possible browser versions presenting vulnerabilities, but showcasing how our solution is valid to launch targeted experiments allowing to identify the presence of vulnerable browsers. Security researchers can use our solution to make a thorough analysis of the presence in the web of browsers with different type of vulnerabilities, even ranking them from most to least problematic vulnerabilities. However, such analysis is out of the scope of this paper.

4) DEMOGRAPHIC TARGETING

The last experiment aims at showcasing the demographic targeting capabilities of our methodology. To this end, we configure a campaign with the following audience parameters: 1) Age: between 18 and 25 years old; 2) Country: Italy; 3) Device: Mobile. We run a 3-days campaign obtaining 13k data samples coming from Android and iOS (note that we got 126 data inputs from Windows Phone, which we consider negligible).

The results first indicate that Android dominates the mobile devices market in the considered demographic group since it accounts for 72% of the data samples, whereas iOS just account with 18%. In the case of browsers, the market share for the considered demographic group is as follows: 1) Chrome Android, 68,40% 2) Safari and Safari WebView, 20% 3) Android browser, 2.5% and 4) Chrome for iOS, 0.8%. These results indicate that young people in Italy prefer Android devices and Chrome browser.

TABLE 4. Number and percentage of impressions for old version usage of the most common browser brands, representing important security vulnerabilities.

Browser	Stable version	Insecure version	Impressions Insecure version
Chrome	62-63	31	13093
Firefox	58-59	30	11849
Safari	11	6	6829

V. CONCLUSION

Measuring the browser marketplace is of interest for companies of different nature, researchers and regulators alike. Existing solutions, based on passive measurement techniques, offer great scalability. However, they present two main drawbacks: First, they require a large monitoring infrastructure, which makes them accessible to just a handful of companies. Second, their passive nature avoids them to offer targeting capabilities.

In this paper, we have presented a novel solution from a technical perspective since it, for the first time, uses active measurements to monitor the browser marketplace enabling targeting capabilities. Moreover, our solution uses the online advertising infrastructure, which nowadays is a commodity used by tens of thousands of companies, as a measurement platform. This democratizes the measurement of the browser marketplace since, contrary to traditional solutions, any interested company or researcher can use it. Finally, our solution offers sufficient scalability to measure the browser marketplace. However, we acknowledge that well-established companies using traditional solutions have reached a larger measurement scale than the one our methodology can achieve at a reasonable cost.

In conclusion, we believe our solution is more suitable for general use by companies and researchers due to its accessibility and targeting capabilities. However, those companies looking for immense scalability should opt for traditional solutions.

REFERENCES

[1] IAB. (2019). *Internet Advertising Revenue Report. 2018 Full Year Results*. Accessed: Oct. 3, 2019. [Online]. Available: https://www.iab.com/wp-content/uploads/2019/05/Full-Year-2018-IAB-Internet-Advertising-Revenue-Report.pdf

[2] Mozilla. (2018). *Security Advisories for Firefox*. Accessed: Oct. 03, 2019. [Online]. Available: https://www.mozilla.org/en-US/security/known-vulnerabilities/firefox/

[3] Chrome. (2018) *Chrome Releases*. Accessed: Oct. 3, 2019. [Online]. Available: https://chromereleases.googleblog.com/

[4] T. Guardian. (2019). *Google Fined 1.49bn by eu for Advertising Violations*. Accessed: Oct. 3, 2019. [Online]. Available: https://www.theguardian.com/technology/2019/mar/20/google-fined-149bn-by-eu-for-advertising-violations

[5] J. Mayer and A. Narayanan. (2019). *Deconstructing Google's Excuses on Tracking Protection*. Accessed: Oct. 3, 2019. [Online]. Available: https://freedom-to-tinker.com/2019/08/23/deconstructing-googles-excuses-on-tracking-protection/

[6] (2019). *W3Counter*. Accessed: Oct. 3, 2019. [Online]. Available: https://www.w3counter.com/globalstats.php

[7] GlobalStats. (2019). *Statcounter*. Accessed: Oct. 3, 2019. [Online]. Available: http://gs.statcounter.com/browser-market-share

[8] S. Frei, T. Duebendorfer, G. Ollmann, and M. May, "Understanding the Web browser threat: Examination of vulnerable online Web browser populations and the 'insecurity iceberg,'" Tech. Rep., Aug. 2008.

[9] T. Duebendorfer and S. Frei, "Web browser security update effectiveness," in *Proc. Int. Workshop Critical Inf. Infrastructures Secur.* Berlin, Germany: Springer, 2009, pp. 124–137.

[10] P. Callejo, C. Kelton, N. Vallina-Rodriguez, R. Cuevas, O. Gasser, C. Kreibich, F. Wohlfart, and À. Cuevas, "Opportunities and challenges of ad-based measurements from the edge of the network," in *Proc. 16th ACM Workshop Hot Topics Netw.*, 2017, pp. 87–93.

[11] (2019). *NetMarketShare*. Accessed: Oct. 3, 2019. [Online]. Available: https://netmarketshare.com

[12] IAB. (2016). *Understanding the programmatic ecosystem: For media buyers & sellers*. Accessed: Oct. 3, 2019. [Online]. Available: https://www.iab.com/guidelines/understanding-programmatic-ecosystem-media-buyers-sellers/

[13] (2019). *AdBLock*. Accessed: Oct. 3, 2019. [Online]. Available: https://adblockplus.org/

[14] (2019). *NoScript*. Accessed: Oct. 3, 2019. [Online]. Available: https://noscript.net/

[15] EUGDPR. (2019). *The eu General Data Protection Regulation*. Accessed: Oct. 3, 2019. [Online]. Available: http://www.eugdpr.org/

PATRICIA CALLEJO received the B.Sc. degree in audiovisual systems engineering and the M.Sc. degree in the field of telematics engineering from the University Carlos III of Madrid, in October 2015. She is currently pursuing the Ph.D. degree in telematics engineering with IMDEA Networks. She was granted by the RIPE Academic Cooperation Initiative (RACI) on RIPE 76 that took place in Marseille, France, in 2018. In the same year, she did an internship at the International Computer Science Institute (ICSI), UC Berkeley, as a part of her Ph.D. She is an author of conferences papers, such as ACM HotNets., ACM CoNEXT, and WWW. She has worked in EU H2020 projects. Her areas of interests include internet measurements, online advertising, privacy, and web transparency.

RUBÉN CUEVAS received the M.Sc. degree in telematics engineering, the M.Sc. degree in telecommunications engineering, and the Ph.D. degree in telematics engineering from the University Carlos III of Madrid, Spain, in 2010, 2007, and 2005, respectively, and the M.Sc. degree in network planning and management from Aalborg University, Denmark, in 2006. In 2012, he was a Courtesy Assistant Professor with the Computer and Information Science Department, University of Oregon. He is currently an Associate Professor and the Secretary of the UC3M-BS Big Data Institute, University Carlos III of Madrid. He has coauthored over 70 papers in prestigious international journals and conferences, such as ACM CoNEXT, WWW, Usenix Security, ACM HotNets, the IEEE Infocom, ACM CHI, IEEE/ACM TON, the IEEE TPDS, CACM, PNAS, Nature Scientific Reports, PlosONE, or Communications of the ACM. He has been the PI of 10 research projects funded by the EU H2020 and FP7 Programs, the National government of Spain and private companies, and in overall participated in 24 research projects. His research in filesharing piracy, online social networks, online advertising fraud, and web transparency has been featured in major international and national media, such as The Financial Times, BBC, The Guardian, The Times, New Scientist, Wired, Corriere della Sera, O' Globo, Le Figaro, El Universal, El Pais, El Mundo, ABC, Cadena Ser, Cadena Cope, TVE, Antena3, and La Sexta. His main research interests include online advertising, web transparency, personalization and privacy, online social networks, and the Internet measurements.

ÁNGEL CUEVAS received the B.Sc. degree in telecommunication engineering, and the M.Sc. and the Ph.D. degrees in telematics engineering from the Universidad Carlos III de Madrid, in 2006, 2007, and 2011, respectively. He is currently a Ramón y Cajal Fellow (tenure-track Assistant Professor) with the Department of Telematic Engineering, Universidad Carlos III de Madrid, and an Adjunct Professor with the Institut Mines-Telecom SudParis. He is a coauthor of more than 50 papers in prestigious international journals and conferences, such as the IEEE/ACM Transactions on Networking, the *ACM Transactions on Sensor Networks*, the *Computer Networks* (Elsevier), the IEEE Network, the *IEEE Communications Magazine*, WWW, ACM CoNEXT, and ACM CHI. His research interests focus on the Internet measurements, web transparency, privacy, and P2P networks. He was a recipient of the Best Paper Award at the ACM MSWiM 2010.