IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received May 5, 2021, accepted May 26, 2021, date of publication May 31, 2021, date of current version June 10, 2021.

Digital Object Identifier 10.1109/ACCESS.2021.3085024

A Measurement Study of Online Tracking and Advertising in Ibero-America

JOSÉ ESTRADA-JIMÉNEZ[1], **JAVIER PARRA-ARNAU**[2,3], **ANA RODRÍGUEZ-HOYOS**[1], **JORDI FORNÉ**[2], **AND ESTEVE PALLARÉS-SEGARRA**[2]

[1]Departamento de Electrónica, Telecomunicaciones y Redes de Información, Escuela Politécnica Nacional (EPN), Quito 170525, Ecuador
[2]Network Engineering Department, Universitat Politécnica de Catalunya (UPC), 08034 Barcelona, Spain
[3]Karlsruhe Institute of Technology, 76131 Karlsruhe, Germany

Corresponding author: Javier Parra-Arnau (javier.parra@upc.edu)

This work was supported by the Spanish Ministerio de Economía y Competitividad (MINECO) through the project ''Secure SMArt Grid using Open Source Intelligence. Data Privacy and Reliable Communications (MAGOS)'', ref. TEC2017-84197-C4-3-R. The research was also supported by the La Caixa Foundation under Grant ID 100010434, in part by the European Union's Horizon 2020 Research and Innovation Programme through the Marie Sklodowska-Curie under Grant 847648, and in part by the Fellowship under Grant LCF/BQ/PR20/11770009.

ABSTRACT The ability of the online marketing industry to track and profile users' Web-browsing activity is what enables effective, tailored-made advertising services. The intrusiveness of these practices and the increasing invasiveness of digital advertising, however, have raised serious concerns regarding user privacy. Although the level of ubiquity of tracking and advertising has been investigated in top-world sites based in North America and Western Europe, the extent to which those practices are carried out in territories with less or no legal coverage —in terms of data protection— has not been studied so far. In this work, we present the first detailed measurement of online tracking and advertising conducted to date in one of those regions, namely, Ibero-America, by analyzing local websites (e.g., education and government sites). In doing so, our measurement study aims to find out how user location as well as the type of publisher may impact on tracking and advertising and thus user privacy. Lastly, our thorough, extensive analysis also explores whether differences are appreciated between Latin America and the EU with regard to the third-party tracking conducted *from* and *towards* the corresponding countries.

INDEX TERMS Online tracking, advertising, privacy risks, Ibero-America.

I. INTRODUCTION

Personalized online advertising is responsible for much of the online tracking performed over users these days. Online advertising platforms are supported by personalization systems that tailor ad-content according to the users' preferences, learned from data collected through Web tracking. Evidently, the more information is gathered, the better the performance of personalization systems, and the higher the profits of the advertising platforms. Since online advertising has become a millionaire business [2] that apparently supports the very existence of the Internet [3], there is a great motivation from multiple agents to collect more and more data, which ultimately exacerbates online tracking.

Online tracking refers to the activity of closely following a user wherever she ''goes'' while browsing the Web. Tracking is possible nowadays as we leave innumerable footprints online, even without noticing it, when browsing. The IP address, operating system, browser type, plugins installed, patches applied, and browsing history are just some examples of the data leaked in a single HTTP request. When multiple browsing sessions are compiled and processed over time, tracking entities can profile and segment users on the basis of, for example, their location [4], shopping habits, sexual preferences [5][1] and political leaning [6]. Obviously, the fact that all such user information may be available to thousands of such entities raises serious privacy and security concerns [7], [8].

In this scenario of ubiquitous tracking —more typical of a dystopian society—, the first potential tracker is the website (also known as publisher) the user visits, as we shall describe later in Section II. Thus, if tracking is carried out by the publisher, it is called *first-party* tracking. In general, the audiences of first parties are pretty segmented, so that

The associate editor coordinating the review of this manuscript and approving it for publication was Claudio Agostino Ardagna.

[1] The cited article explores tracking and privacy risks on pornography websites. The analysis of over twenty thousands of such websites showed that 93% leak user data to a third-party.

the tracking those parties might perform is usually innocuous. Some exceptions are the ecosystems represented by the Internet giants (e.g., Facebook, Google), which concentrate services for millions of users within a single corporation.

In addition to first-party tracking, a single user web-request commonly triggers connections from their browser to several *third parties* that receive part of the aforementioned information. This information is used by third-parties to support real-time services such as personalized advertising, media hosting (by content distribution networks), load balancing and social networks. Figure 1 illustrates the interactions triggered by a browser request, which enable first and third-party tracking. Figure 2 shows, on the other hand, the large number of connections to third parties (information flows) that may occur when a user visits three websites.

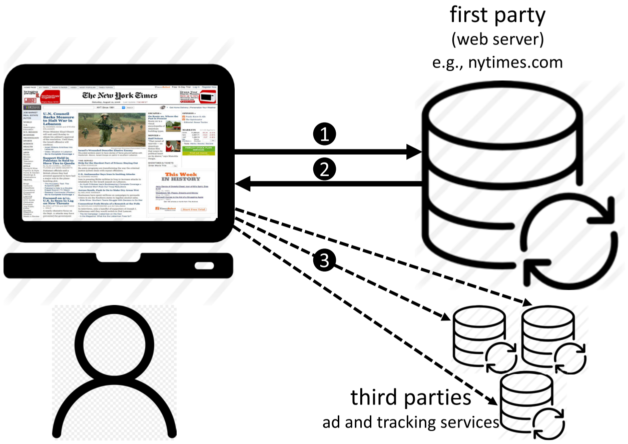


FIGURE 1. **A single user's visit to a website (1) generates a response (2) that commonly includes redirection commands that spawn further connections to third parties (3).**

Unlike physical user tracking, e.g., in a street, online tracking is much less evident for users because it is carried out on the background of web browsing. Unfortunately, there is very little evidence available for users to realize the latent pervasiveness of such practice.

Due to its prevalence on the Web, measuring online tracking could be a way to characterize the privacy risks of Internet users. The severity of those risks may be illustrated through different indicators such as the level of exposition of user interactions to third parties, the concentration of user information on a few advertising companies, the dynamic behavior of tracking for websites belonging to certain categories, or the suspicious requests to third parties triggered when accessing government web sites [10].

While related work [11] has investigated tracking and advertising in the most popular sites worldwide,[2] our measurement study focuses on the online tracking and advertising triggered by *local websites* (e.g., education and government sites), specifically those from *Ibero-American*[3] countries. In doing so, our experimental analysis aims at finding out how user location as well as the type of publisher may impact on the tracking and advertising interactions and thus user privacy. Therefore, unlike previous research, our study aims to investigate a very heterogeneous region like Latin America (LATAM) and compare the tracking and advertising practices with those in the European Union (EU), in terms of the adoption of modern personalization systems —including behavioral advertising— and the maturity of user perceptions regarding privacy. In this same regard, our analysis will also explore whether differences might be appreciated between LATAM and the EU with regard to the third-party tracking conducted *from* and *towards* the corresponding countries. This might be of particular interest since the General Data Protection Regulation (GDPR)[4] became enforceable in May 2018 in the EU and the European Economic Area, while LATAM is considerably less effective in this field —let alone many of the countries in the latter region lack regulation for privacy.

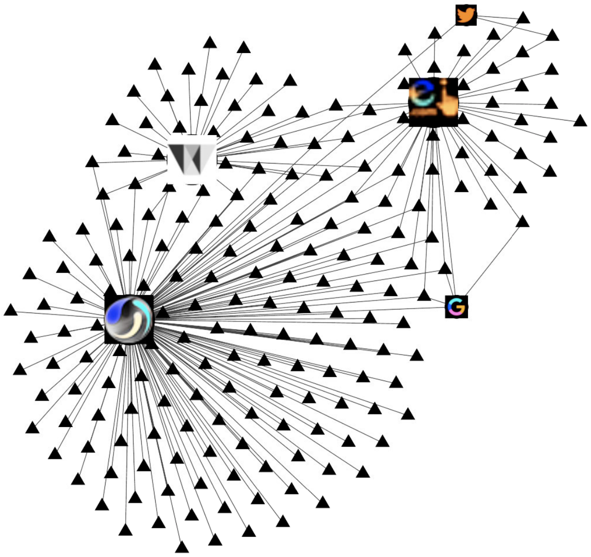


FIGURE 2. **Illustration of the multiple connections to third parties (more than 50) generated in the background after visiting only 3 sites. The points where connections originate represent the websites while the little triangles represent the third parties contacted. This figure was obtained through the browser extension Disconnect [9].**

The rest of this paper is organized as follows. Section III describes the methodology followed in this work, including data collection, processing, and experiments. Section IV presents our measurement results regarding online tracking and advertising in Ibero-America. Finally, conclusions are drawn in Sec. V.

[2] The cited work restricts to top-world sites based in North America and Western Europe.

[3] Ibero-America is a region in the Americas comprising countries or territories where Spanish or Portuguese are predominant languages, usually former territories of Portugal or Spain.

[4] The GDPR is considered the toughest privacy and security law in the world.

II. BACKGROUND

A. ONLINE TRACKING AND ADVERTISING

This section examines the online advertising ecosystem, providing the reader with the necessary depth to understand the technical contributions of this work. For a detailed, complete explanation on the subject, the reader is referred to [12], [13].

B. KEY ACTORS

The online advertising industry is composed by a considerable number of entities with very specific and complementary roles, whose ultimate aim is to display ads on Web sites. Publishers, advertisers, ad platforms, ad agencies, aggregators and optimizers are some of the parties involved in the ad-delivery process [14]. Despite the enormous complexity[5] and constant evolution of the advertising ecosystem, the process whereby ads are presented on Web sites is usually characterized or modeled in terms of publishers, advertisers and ad platforms [16]–[22]. Next, we provide a description of these three key actors:

- A *publisher* is an entity that owns a Web page (or a Web site) and that, in exchange of some economic compensation, is willing to place ads of other parties in some spaces of its page (or site). An example of publisher is The New York Times' Web site.
- An *advertiser* is an entity that wants to display ads on one of the spaces offered by a publisher, and is disposed to pay for it. Advertisers typically engage the services of one or several *ad platforms* (described below), which are the ones responsible for displaying their ads on the publishers' sites. As we shall explain later in Section III, there exist two ad-platform models, allowing users to have two different roles. In the traditional albeit prevailing approach, advertisers indicate the targeting objective/s most suitable for their campaigns, that is, to which users they want their ads to be shown. For example, an advertiser may want the ad platform to serve its ads to an audience interested in politics or to people living in France. Advertisers must also specify the amount of money they are willing to pay each time their ads are displayed, and each time users click on them.[6] On the contrary, in the recently established model of *real-time bidding* (RTB), ad platforms allow advertisers to bid for each ad-impression at the time the user's browser loads a page. This model enables advertisers to make their own decisions rather than relying on an intermediary to make decisions for them [13].
- An *advertising platform* or *ad platform* is a group of entities that connects advertisers to publishers, i.e., it receives ads from advertisers and places them on the spaces available at publishers. To this end, ad platforms track and profile users with the aim of targeting ads to their interests, location and other personal data. As we shall describe in greater detail in the next subsection, traditional ad platforms carry out this targeting on their own, in accordance with the campaign requirements and objectives specified by advertisers. RTB-based ad platforms, on the other hand, share certain user-tracking data with advertisers, which then take charge of selecting who suits them by deciding which user to bid for. Some examples of ad platforms include DoubleClick, Gemini and Bing Ads, owned respectively by Google, Yahoo! and Microsoft.

III. EXPERIMENTAL METHODOLOGY

We aim to study the impact of online tracking and advertising in Ibero-America by measuring the third-party traffic triggered when browsing websites in this region. In the next subsections, we shall describe how the data related to this traffic was generated, collected and processed. It is important to remark that our measurement study relies on the `ads.txt` standard, increasingly being adopted by websites for some years now. `ads.txt` is a project promoted by the Internet Advertising Bureau (IAB) to increase transparency in the programmatic advertising ecosystem and prevent fraud. It encourages publishers to publicly inform the companies they have authorized to sell their advertising inventory (ad spaces). Such publication is done through a text file (so much like the `robots.txt` standard) called `ads.txt` in the root context of the website.

A. SELECTION AND CATEGORIZATION OF WEBSITES

1) SELECTION OF WEBSITES

We started by selecting the countries of Ibero-America whose websites will be analyzed. We chose the countries in this region that allowed us a VPN connection so that web traffic could be generated from such locations. This included Spain and Portugal from the EU and the following LATAM countries: Argentina, Brazil, Chile, Colombia, Costa Rica, Ecuador, Mexico, Peru, Uruguay, and Venezuela. For some tests involving web traffic directed *to* LATAM, we also included other countries.

Next, we gathered the most popular websites within each of the aforementioned countries. For this, we manually selected only the local websites from the top 500 ranking published for each country by the Alexa Top Sites service offered by Amazon [23]. For other experiments, we also collected the top 500 global websites also using this service as source. In a nutshell, our experimental scenario consisted of 12 countries, 2 076 Ibero-American websites and 500 top-world websites.

2) CATEGORIZATION OF WEBSITES

With the aim of understanding the influence of the content published in websites on the phenomena studied in this work, we also labeled each website tested according its content. We *manually* categorized each of the websites, with the support of the Site Safety Center tool of Trend Micro [24] and the categorization software developed by [25].

[5] The intricacy of the advertising ecosystem is often illustrated in conferences and related venues with the diagram available at [15].

[6] In the terminology of online advertising, these quantities are referred to as the cost-per-impression (CPI) and the cost-per-click (CPC), respectively.

B. DATA COLLECTION

The data we studied mainly included all the third-party requests spawned by visiting a website. Assessing the magnitude of such traffic, the destination third-parties participating, and even the tracking information they may set in the user side (cookies), may help to unveil the inherent privacy risks of users browsing these sites.

Our experiments involved simulating websites visits to trigger third-party traffic. For the thousands of websites in our scenario, we performed this (including data collection) automatically through OpenWPM, a versatile tool used for web measurement [11]. OpenWPM offers a programmable interface to orchestrate the main functions of a web browser, thus allowing automated web crawling and collection of tracking-related information (redirects, cookies and third-party calls) that is stored in a SQLite database. To study the adoption of the `ads.txt` advertising standard, we also employed this tool.

C. EXPERIMENTS

We performed several tests simulating web traffic from different countries in Ibero-America, given our interest in studying potential privacy risks in this region. As commented previously, we utilized a VPN service to connect to each of said countries and send web traffic to local websites.

First, we generated *local traffic*, i.e., visits from each country to websites in the same country. Furthermore, we simulated visits from (a country in) EU to websites in LATAM countries. This way we tried to detect different effects when the same publishers are visited from diverse locations. Since the particular website being visited is likely to be determinant (especially if widely popular),we also simulated web traffic from Ibero-American countries to top-world sites.

Through OpenWPM, we collected information on (1) the third-party requests triggered by simulated user browsing, and on (2) added third-party cookies set on the user side. From said data, we also measured advertising related requests, which, associated with a personalized service, may imply higher privacy risks. To measure the dynamics of advertising interactions, we classified third-party requests as ad related or not, by resorting to available libraries that, based on ad blocking lists [26], facilitate the detection of such type of traffic.

We aimed at unveiling privacy risks by measuring the intensity of third-party traffic (including ad related interactions) and counting the number of entities behind such traffic, and the cookies set in the browser. Third-party traffic, measured as the mean number of third party requests triggered from websites, is a first approach to unveil potential privacy issues. The greater the indirect and non-consensual traffic from users, the more user information is released through such flows. In the same line, we obtained the number of third-parties receiving said traffic as a proxy to the number of potential third-party trackers. We also counted the mean number of cookies set by third-parties, and particularly those cookies more likely to be related to user pseudo-identifiers. Since online tracking tightly depends on this information, measuring this parameter contributed to our objective of detecting privacy risks.

To ascertain the impact of the type of content served by publishers on the intensity of third-party tracking, we tabulated the information collected according to the categories of websites. Since the consumers of certain (categories of) content might be more relevant in economic terms, the sites associated with such content might be more exposed to third-party tracking and privacy risks.

Finally, we identified the Ibero-American publishers where the `ads.txt` file was present. From the data obtained, we studied the adoption of this standard in Ibero-America and the monopolistic influence of some companies on the advertising ecosystem.

IV. EXPERIMENTAL RESULTS

A. THIRD-PARTY TRAFFIC

In Fig. 3, we illustrate the impact of third-party traffic triggered by browsing Ibero-American web sites locally, i.e., from each country. Such map representation enables us to illustrate the geography, size, and potential population of involved countries. This perspective reveals the marked heterogeneity not only among Latin American countries, but particularly among them and EU countries (here represented by Spain and Portugal).

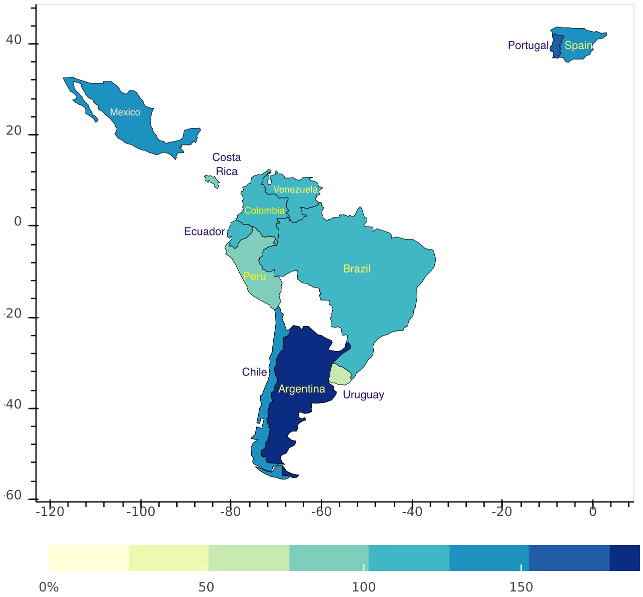


FIGURE 3. Heat map illustrating the mean number of third-party requests triggered from local traffic in Ibero-American countries. The countries included are those to which we could connect through a VPN service to simulate web traffic.

Much traffic towards third-parties is observed in this context. Besides the great number of third-party requests triggered in general by a single visit to publishers, we evidence that such traffic is not homogeneous along countries. Since the volume of interaction with third-parties could be a proxy for potential privacy risks, the resulting impact on privacy would also vary from country to country.

Having a greater volume of third-party traffic, it stands out that some large and populated countries, such as Brazil and Peru, receive a lower impact than that in Portugal and Spain, although the latter are much smaller and less populated. On the other hand, Argentina, Mexico, and Chile do show an important number of third-party requests spawned by local traffic.

Beyond the disparity in Latin American countries, European publishers and users seem to be more attractive targets for third-party tracking. This might imply a higher risk for such users since more entities and more personal information would be involved. However, many of these third-party requests could have the same destination, so it is convenient to identify the recipient entities (third-party trackers) by filtering the domain names from their destination URLs. As noted above, insofar these entities receive so much indirect user traffic, they might become attackers, not only against user privacy but also against user security.

As commented in Section I, the number of third-party domains behind online tracking may better reflect a potential privacy attack scenario since it could serve as a proxy of the number of third-party entities collecting information (third-party trackers). As depicted in Fig. 4, hundreds of entities were found receiving third-party requests, indirectly, from users, when locally visiting Ibero-American sites. European countries had a very similar number of potential third-party trackers while in Latin America the situation is less uniform. In Brazil and Mexico, by far the most populated countries, we found the greatest number of third-party entities. However, we think the difference with other countries is certainly minor, considering the variation in population.

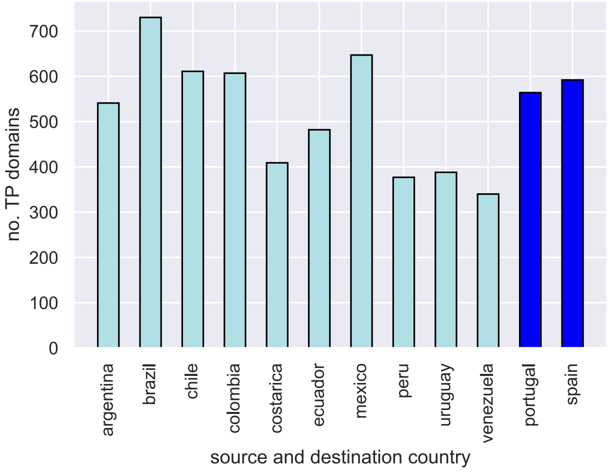


FIGURE 4. Number of third-party domains contacted as a result of local traffic within Ibero-American countries.

One might think that websites in a country of hundreds of millions of citizens, such as Brazil, would spawn much more third-party traffic than those in a country with a couple of tens of millions (e.g., Chile, Spain, and particularly Portugal), but the difference is not as marked as the population. There might be some reasons. First, it could reflect a preference of potential trackers to certain type of population. However, third-party requests are also generated by technology services, e.g., content distribution, commonly used by publishers. Thus, more entities behind third-party traffic could also indicate more tech supporting websites. In any case, there would be more privacy risk for a user if he is targeted in a small population than in a big one.

By weighting the number of third-party domains by the population of countries (in millions of inhabitants) as shown in Fig. 5, we tried to capture the latter effect. We can see in this figure that, despite being very sparsely populated, Costa Rica, Uruguay, and Portugal spawn a lot of third-party traffic. Despite its size, these are relative rich countries, in particular when compared with the average in Latin America, which we think could explain this behavior to some extent. Paradoxically, richer countries are more prone to implementing strong privacy legislation, which in this context does not seem to discourage third-party traffic. Note that this behavior is measured when crawling local sites of each country *from* the same country.

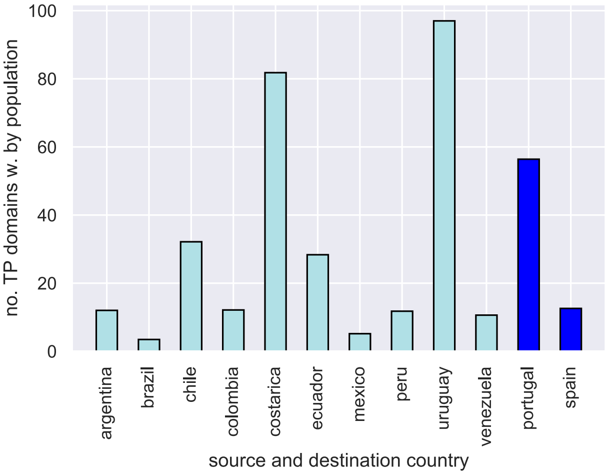


FIGURE 5. Number of third-party domains contacted as a result of local traffic, weighted by population, within Ibero-American countries.

In our attempt to identify third-party tracking, and specifically personal data leaking, another interesting indicator might be the mean number of third-party cookies set in the user's browser. Recall from Section II that third-party cookies may be used to transport user identifiers that online trackers employ to recognize a user when she visits a publisher. This way, a tracker is able to "follow" users and associate information to their profiles. We cataloged these cookies as tracking cookies or identifying cookies (ID cookies) if their lengths were greater than 6.

When processing the data obtained from local traffic within Ibero-American countries, we found that a user browsing local sites from Portugal would receive 14 ID cookies

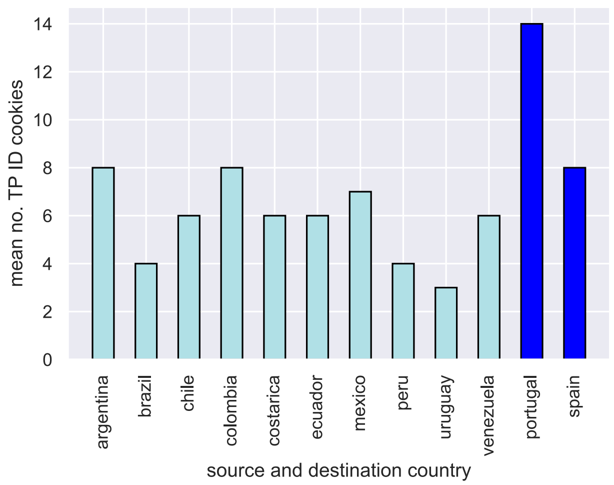


FIGURE 6. Mean number of identifying third-party cookies found as result of local traffic within Ibero-American countries.

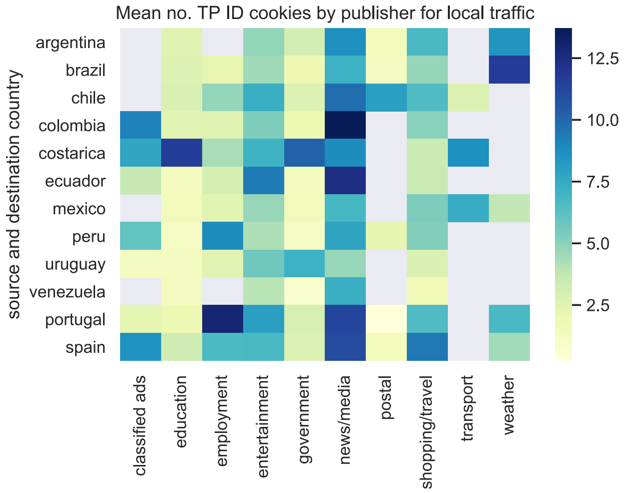


FIGURE 7. Mean number of ID third-party cookies set from local traffic within Ibero-American countries, organized by category of publisher.

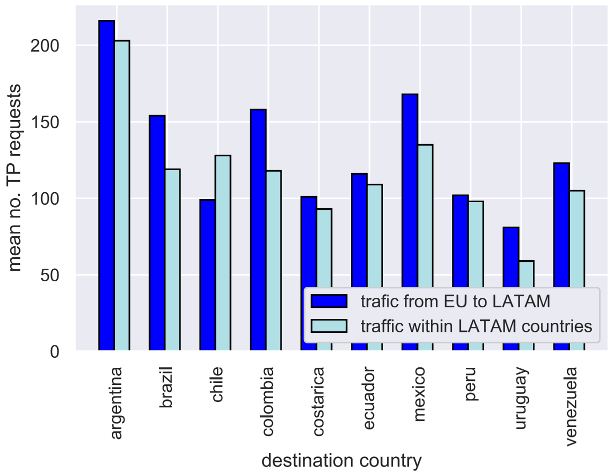


FIGURE 8. Mean number of third-party requests triggered by web traffic from the EU to LATAM.

on average, as depicted in Fig. 6. Spain, Argentina, and Colombia follow with 8 ID cookies on average. This figure shows that this number varies along the different countries but suggests a great interest of Portuguese local sites in tracking local users.

Since we categorized each of the publishers visited along our experiments, we could represent in Fig. 7 the potential influence of the content delivered by publishers on the tracking performed over users. As shown in other works [11], [27], `news/media`, `entertainment`, and `shopping/travel` are the categories concentrating more third-party tracking, in this case when local traffic is studied in Ibero-American countries. Some particularities should be noted about some countries: users from Costa Rica would be receiving a lot of third-party tracking triggered by education and government sites; and, from Brazil, would occur the same with `weather`.

As a consequence, some of these categories may entail more privacy risks than others. For instance, `education` could involve audiences including children or teenagers who are clearly more vulnerable to abuse, more if the sites belonging to these category require any kind of login. Also, probably, government sites should not be tracking their citizens, even worse to ''profit'' from their interactions if coupled, e.g. with advertising platforms.

When comparing traffic to LATAM originated locally vs. from the EU, traffic from the EU spawns, in general, more third-party requests per website as shown in Fig. 8. Namely, when traffic originates in EU, more third-party interactions are observed in most of LATAM countries. Chile is the exception.

Although the intensity of third-party traffic is higher for web browsing from the EU than internally in LATAM, the total number of domains or entities behind third-party traffic is very similar, no matter where the web traffic originates, as depicted in Fig. 9. The same entities would be involved in third-party traffic but changing the tracking strategy depending on the location of the user. However, when comparing web traffic from the EU vs web traffic from LATAM (to the same LATAM websites), the total number of third-party entities does not vary significantly, although web traffic from the EU would be covered by the GDPR. It seems, then, that the extraterritorial effect of such a regulation is not effective in this context.

It is still compelling that very small countries such as Uruguay and Costa Rica might be tracked by a lot of entities despite their small population. When weighting such number of third-party domains by the population of countries, as illustrated in Fig. 10, this detail is evidenced. On the other hand, note how huge and highly populated countries such as Brazil and Mexico show an opposite phenomenon: a relatively small number of third-party entities. This again characterizes a

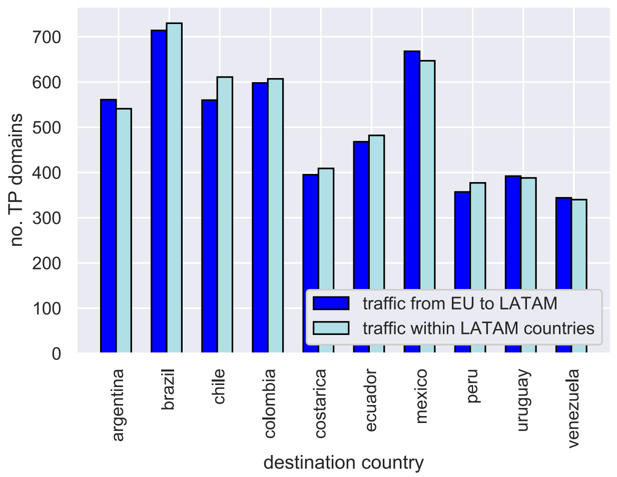


FIGURE 9. Total number of third-party domains found behind web traffic from the EU to LATAM.

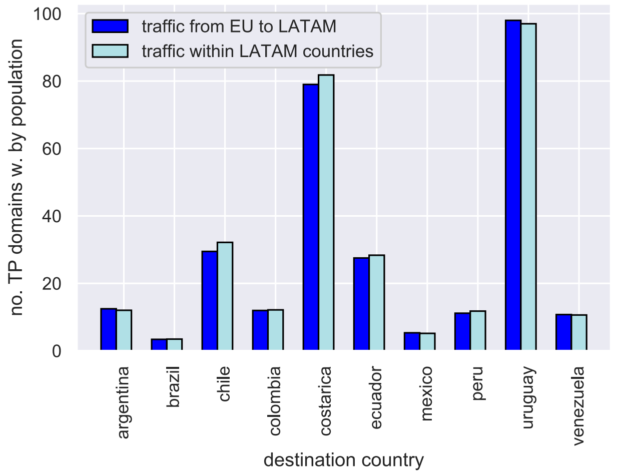


FIGURE 10. Total number of third-party domains found behind web traffic from the EU to LATAM.

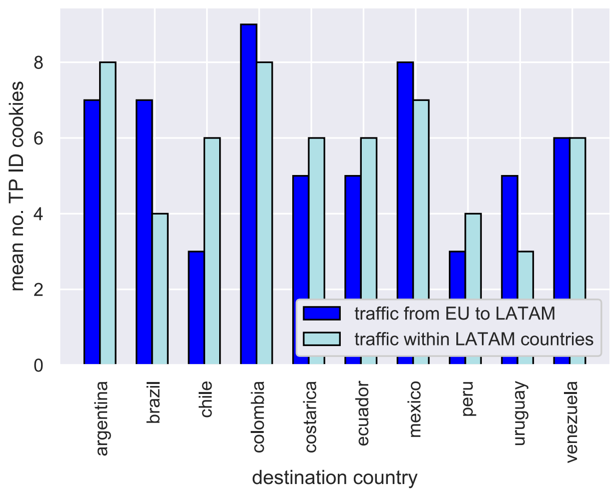


FIGURE 11. Mean number of third-party ID cookies found behind web traffic from the EU to LATAM.

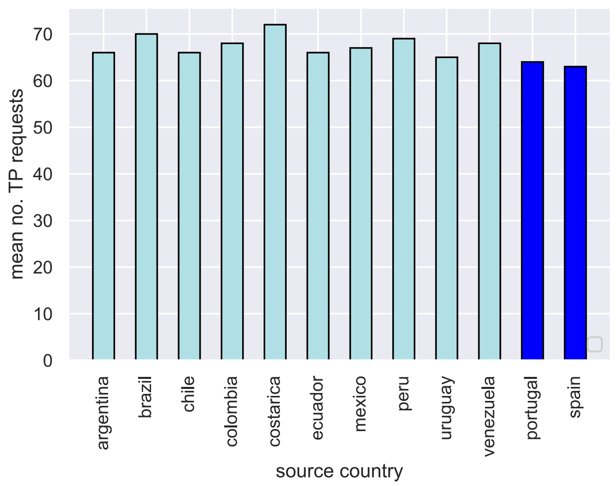


FIGURE 12. Mean number of third-party requests triggered by web traffic to top-world sites.

marked heterogeneity in the Latin American context, where tiny populations might become the target of several external entities. Naturally, individuals in such groups would be more exposed to privacy risks than those in larger groups.

For some countries, the number of identifying cookies is slightly greater when web requests are generated locally (Argentina, Costa Rica, Ecuador, Peru) than when coming from the EU, although in Chile such number is doubled (see Fig. 11). The opposite occurs in the rest of the countries (Brazil, Colombia, Mexico, Uruguay). The inherent potential tracking, then, varies from country to country and apparently regardless its origin is EU or LATAM. Probably, in this regard, a more individualized study should be developed to unveil the reasons of this behavior.

When testing web traffic from Ibero-American countries *to the top-world sites*, we found out in Fig. 12 that the number of derived third-party requests seems to be, in general, stable along all the countries analyzed. Interestingly, the same effect is shown when illustrating in Fig. 13 the total number of third-party domains behind such interactions. Note, however, that this number of entities is almost ten times greater than that when the web traffic was directed to local sites.

The behavior of third-party tracking may be significantly more intense when visiting globally popular sites, no matter the source of such visits. Thus, in this scenario, the relative impact on small countries such us Costa Rica, Uruguay and Portugal might be greater given its small population.

The mean number of identifying cookies set when visiting top-world sites from Ibero-America is also homogeneous along the countries tested. This number is certainly greater than when web traffic is directed to LATAM countries, despite the number of entities behind is ten times higher. That the number of potential tracking cookies does not grow

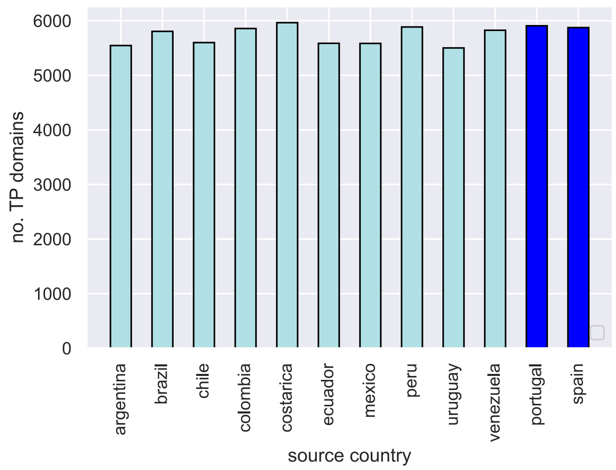


FIGURE 13. Total number of third-party domains found behind web traffic to top-world sites.

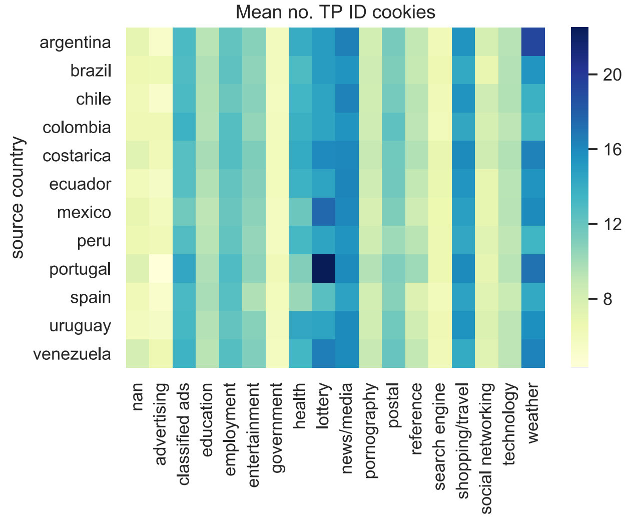


FIGURE 14. Mean number of ID third-party cookies set by web traffic to top-world sites, from Ibero-American countries, organized by category of publisher.

as significantly as the number of third-party domains may suggest that widely popular websites could be also resorting to other more sophisticated tracking mechanisms.

The mean number of identifying cookies set when visiting top-world sites from Ibero-America is also homogeneous along the countries tested. This number is certainly greater than when web traffic is directed to LATAM countries, despite the number of entities behind is ten times higher. That the number of potential tracking cookies does not grow as significantly as the number of third-party domains may suggest that widely popular websites could be also resorting to other more sophisticated tracking mechanisms.

Finally, as shown in Fig. 14, `news/media` and `shopping/travel` were again the categories spawning more third-party tracking when browsing top-world sites. Since a wider spectrum of sites was covered in this experiment, other categories were also relevant in terms of potential tracking such as `weather`, `lottery` and `health`. Regarding privacy risks, note how, in this context, sites potentially serving, collecting and even sharing very sensitive information (health or pornography) could be the source of intense third-party tracking.

In any case, gateways of information such as digital newspapers or sites involving any kind of commerce are those spawning a great deal of tracking, either because of their big audiences or because user willingness to buy in these sites is high, no matter where users are located.

B. ONLINE ADVERTISING

To study the impact of online advertising in the context of Ibero-America, we resort to measuring ad related third-party tracking, and to gathering the information hosted by publishers in compliance with the `ads.txt` standard.

1) ANALYSIS THROUGH THE `ads.txt` STANDARD

Recall that the `ads.txt` standard promotes that websites publicly inform the companies or domains they have authorized to sell their advertising inventory through a text file called `ads.txt` in the root context of the website. The publication of this file by a website shows a strong commitment of the publisher with the advertising industry and could give

In Fig. 15, the adoption of the standard `ads.txt` is depicted. Longer curves such as those of Portugal and Spain indicate that more websites host this file including the ad exchanges authorized to sell ad spaces from said websites. The figure shows that the adoption of `ads.txt` in Latin American countries is still reduced compared to the countries analyzed in the EU. Only Brazil and Argentina present similar levels of adoption. In any case, the number of records per website goes from 1 to 150 for most countries. The specific domains involved are studied below.

To show how third-party entities are distributed along websites in Ibero-America, based on the records found in the `ads.txt` files, we first plotted Fig. 16. This ECDF shows that 1% of the third-party domains found appear on *more* than 20% of the websites crawled (more than 70% of the sites hosting an `ads.txt`). Thus, the concentration of advertising in a few entities is evident in this context; to give an example, `google.com` were engaged with all the sites that had adopted this standard, and 4 others(`rubicon, appnexus, openx and pubmatic`) were engaged with more than 70% of the sites that had an `ads.txt` file.

When analyzing websites by category in Table 1, we found that `ads.txt` is widely adopted by `news/media` sites, followed by `entertainment` and `shopping/travel`, as it happened with third-party tracking. Besides, in Table 2, we depict the mean number of third-party domains identified within `ads.txt` files. Although, a few `education` related publishers were found that have adopted this standard, the records in the `ads.txt` file in these sites included even

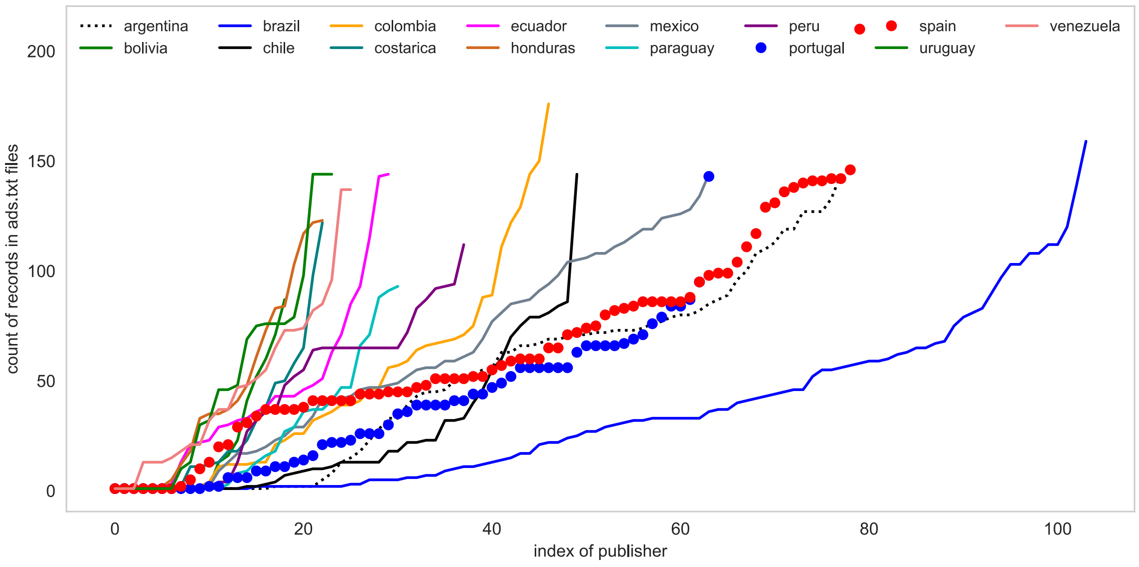


FIGURE 15. Number of records in `ads.txt` files found in websites of Ibero-American countries.

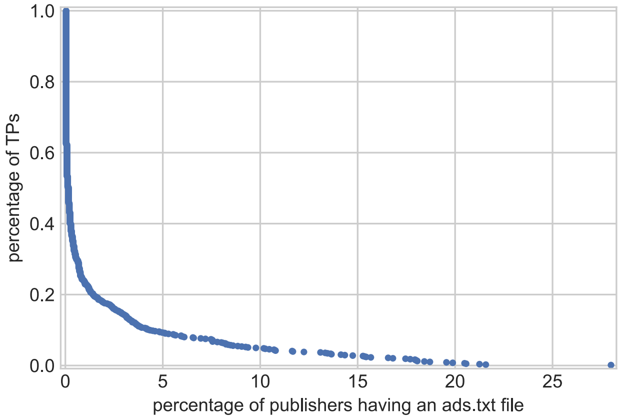


FIGURE 16. ECDF % of web sites covered by third-party domains in Ibero-American countries.

more third-party domains than `news/media`. Note that no government hosted an `ads.txt` file in this scenario.

2) ANALYSIS THROUGH THE AD RELATED TRAFFIC

When analyzing ad related traffic generated in Latin American sites, as a result of browsing from LATAM and EU, we can see that, in general, web traffic from the EU triggers a little more ad related tracking than traffic from LATAM, except in Chile and Peru. This was illustrated in Fig. 17 where the mean number of ad related requests spawned by country is presented. Websites from Argentina trigger, by far, the highest intensity of ad related traffic, followed by Venezuela. The context is certainly heterogenous but, as stated in previous paragraphs, the source of web traffic might be influencing the ad related tracking derived when browsing LATAM local sites.

TABLE 1. Number of websites having adopted the `ads.txt` standard in Ibero-American countries by category.

category	no. of websites with `ads.txt`
`news/media`	337
`entertainment`	138
`shopping/travel`	105
`employment`	29
`education`	13
`classified ads`	12
`search engine`	8
`weather`	7
`lottery`	2
`legal`	2
`postal`	1
`health`	1

TABLE 2. Mean number of third-party domains found in `ads.txt` files by category of website in Ibero-American countries.

category	mean no. of third-party domains in `ads.txt`
`health`	82
`education`	61.3
`news/media`	57.69
`search engine`	49.5
`entertainment`	47.42
`lottery`	33.5
`classified ads`	24.17
`employment`	22.89
`weather`	22.85
`shopping/travel`	15.4
`legal`	11.5
`postal`	1

We can see that traffic from the EU to LATAM would trigger more ad related tracking than traffic to LATAM. Thus, the location of users in this scenario would affect the level of user tracking. Furthermore, if user browsing originates in LATAM and directs to top-world sites, this

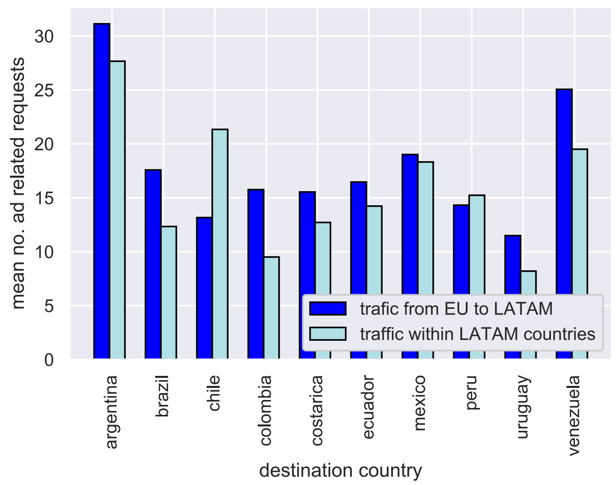


FIGURE 17. Mean number of ad related requests spawned from web traffic to LATAM when originated locally (LATAM) and from the EU.

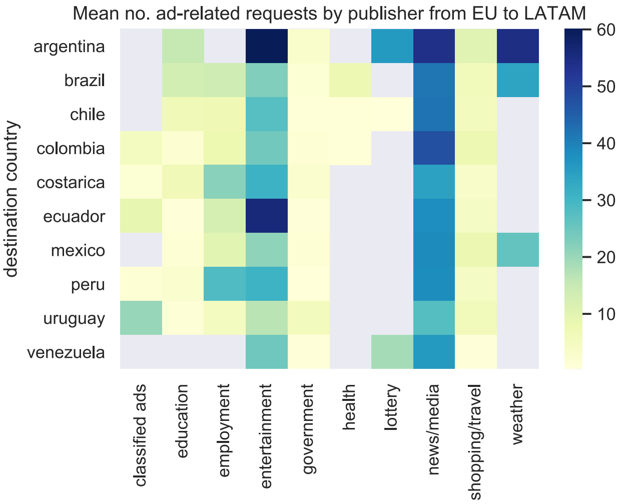


FIGURE 19. Mean number of ad related requests triggered by web traffic from the EU to LATAM.

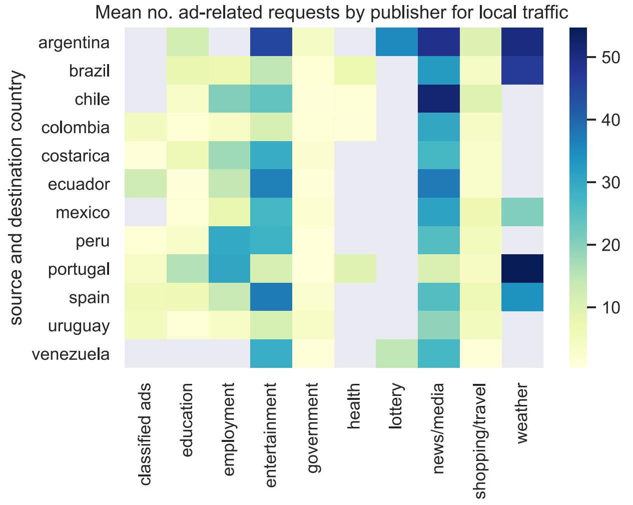


FIGURE 18. Mean number of ad related requests triggered by local traffic.

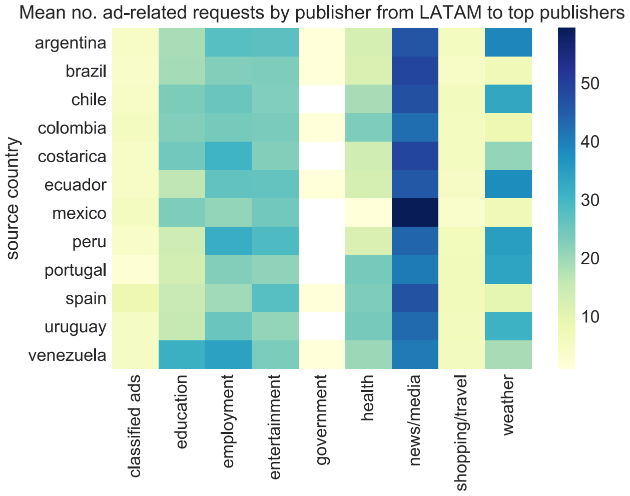


FIGURE 20. Mean number of ad related requests triggered by web traffic from top-world sites.

type of tracking increases notably; its proportion with respect to all third-party traffic is, however, reduced. Consequently, we can confirm, although not entirely unexpectedly, that users browsing very popular websites would be more exposed to ad related tracking.

Based on the content served by publishers, we found that categories `news/media` and `entertainment` exhibited the highest ad related traffic both when web traffic comes from LATAM and from the EU. For some countries, `employment` and `weather` caused important ad related traffic. This is depicted in Figs. 18 and 19.

Moreover, the impact of this third-party traffic in `government` sites is minor, but it is present in all countries, although we believe commercial advertising for profit should not be engaged with public sites already funded by the taxes of citizens. We found out 103 and 135 government sites triggering ad related tracking from visiting LATAM sites locally and from the EU, respectively. More than 90% of these government sites were covered by third-party domains coupled with Google.

When measuring ad related requests triggered by web traffic to top-world publishers, `news/media` sites showed an important ad tracking activity, followed by `entertainment`, as with the contexts analyzed previously. This is shown in 20. However, in these very popular sites, a significant tracking activity is also observed for several other categories such as `employment`, `education` and `health`. The intrinsic privacy risks involved in this interactions, thus, might be present in a wider spectrum of publishers, including some categorized as `education` and `health` websites, which could be collecting, processing and sharing very sensitive data.

We calculated the prevalence of these third-party entities along local websites in Ibero-America, and presented the data

TABLE 3. Prevalence of third-party domains behind ad related traffic along websites visited in Ibero-American countries *locally*.

third-party domain	% of websites covered
doubleclick.net	42.41
googlesyndication.com	31.27
google.com	27.63
ampproject.org	17.54
pubmatic.com	16.38
2mdn.net	12.06
google.com.co	9.00
google.com.br	7.67
yahoo.com	7.38
google.com.ar	5.88

TABLE 4. Prevalence of third-party domains behind ad related traffic along websites visited in Ibero-American countries *from the EU*.

third-party domain	% of websites covered
google.es	43.27
doubleclick.net	35.71
googlesyndication.com	25.79
google.com	22.67
2mdn.net	15.40
ampproject.org	13.56
pubmatic.com	13.38
yahoo.com	5.42
googleadservices.com	4.09
krxd.net	3.98

in Tables 3 and 4. We found that domains owned by Google were massively behind advertising traffic. In particular among the 10 most prevalent third-party domains, at least half of them belonged to this company. `doubleclick.net`, for example, appeared at more than 40% of the websites crawled from LATAM, a similar coverage of `google.es` when web traffic originated in EU. Besides, the third-party domains involved and their prevalence vary if traffic to LATAM originates in EU; the prevalence of some third-party domains along websites is, in fact, slightly higher. Again, this would suggest that users browsing from the EU would be more exposed to the inherent user profiling and tracking of online advertising than LATAM users.

V. CONCLUSION

User location affects the intensity of third-party traffic, including advertising related flows of information, triggered from browsing Ibero-American websites. This is evident first when assessing local web traffic, i.e., when browsing these sites from their country of origin. LATAM and EU countries in this region showed a significant heterogeneity among them in this respect. But the influence of location is also shown when measuring the interactions spawned by web traffic from the EU to LATAM. We found out that traffic originating in EU spawns more third-party interactions. Despite stricter privacy regulations in the EU, users from such locations would be more exposed than users from LATAM to the targeting and profiling performed by external trackers.

Interestingly, the total number of potential trackers (third-party domains) found behind third-party requests, including ad related requests, is similar either web traffic is local or from the EU. It seems the same entities are involved in third-party traffic but change their tracking strategy depending on the location of the user.

Users from particular locations might be more "relevant" for online tracking and advertising. Despite the very small population of some countries, the level of third-party traffic derived from them is comparable to that of other huge countries. Portugal is an example.

When top-world sites are the destination of web traffic generated from Ibero-America, the user location parameter is less relevant to the intensity of potential third-party tracking, since the mean number of requests triggered is similar among the source Ibero-American countries. Privacy risks arise higher in this context since more third-party requests are spawned and significantly more third-party domains are found behind.

Websites or publishers whose content falls into the categories `news/media`, `entertainment` or `shopping/travel` showed the greatest levels of potential third-party tracking and third-party entities. Thus, the interactions with said websites might entail more privacy risks in different contexts. Note that when web traffic goes to top-world websites, this risk is extended to other categories related to the collection of sensitive data, such as `health` or `education`.

The level of adoption of the `ads.txt` standard in LATAM is still low compared to that of the EU. The information it reveals is valuable to study the third-parties (ad exchanges) officially coupled with publishers. It evidences the concentration of advertising in a single company, Google; and in publishers associated with the categories `news/media`, `entertainment` or `shopping/travel`.

ACKNOWLEDGMENT

This research work is part of J. Estrada-Jiménez's Ph.D. thesis [1].

REFERENCES

[1] J. Estrada-Jiménez, "Privacy in online advertising platforms," Ph.D. dissertation, Dept. Netw. Eng., Universitat Politècnica de Catalunya, Barcelona, Spain, Oct. 2020.

[2] M. Graham. (2018). *Digital ad Revenue in the US Surpassed $100 Billion for the First Time in 2018*. [Online]. Available: https://www.cnbc.com/2019/05/07/digital-ad-revenue-in-the-us-topped-100-billion-for-the-first-time.html

[3] C. Gayomali. (2014). *It Would Cost Each User $232 a Year for an Ad-Free Internet, Study Finds*. [Online]. Available: https://www.fastcompany.com/3034670/it-would-cost-each-user-232-a-year-for-an-ad-free-internet-study-finds

[4] A. Rodriguez-Carrion, D. Rebollo-Monedero, J. Forné, C. Campo, C. Garcia-Rubio, J. Parra-Arnau, and S. Das, "Entropy-based privacy against profiling of user mobility," *Entropy*, vol. 17, no. 6, pp. 3913–3946, Jun. 2015.

[5] E. Maris, T. Libert, and J. R. Henrichsen, "Tracking sex: The implications of widespread sexual data leakage and tracking on porn Websites," *New Media Soc.*, vol. 22, no. 11, pp. 2018–2038, Nov. 2020.

[6] N. Confessore. (Apr. 2018). *Cambridge Analytica and Facebook: The Scandal and the Fallout so Far*. Accessed: Apr. 23, 2021. [Online]. Available: https://www.nytimes.com/2018/04/04/us/politics/cambridge-analytica-scandal-fallout.html

[7] M. Ruckenstein and J. Granroth, "Algorithms, advertising and the intimacy of surveillance," *J. Cultural Economy*, vol. 13, no. 1, pp. 12–24, Jan. 2020, doi: 10.1080/17530350.2019.1574866.

[8] K. D. Martin and R. W. Palmatier, "Data privacy in retail: Navigating tensions and directing future research," *J. Retailing*, vol. 96, no. 4, pp. 449–457, Dec. 2020. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0022435920300658

[9] Disconnect.me. *Private Browsing*. Accessed: Dec. 16, 2015. [Online]. Available: https://disconnect.me/disconnect

[10] J. Estrada-Jiménez, J. Parra-Arnau, A. Rodríguez-Hoyos, and J. Forné, "Online advertising: Analysis of privacy threats and protection approaches," *Comput. Commun.*, vol. 100, pp. 32–51, Mar. 2017.

[11] S. Englehardt and A. Narayanan, "Online tracking: A 1-million-site measurement and analysis," in *Proc. ACM SIGSAC Conf. Comput. Commun. Secur.*, Oct. 2016, pp. 1388–1401.

[12] K. Varnali, "Online behavioral advertising: An integrative review," *J. Marketing Commun.*, vol. 27, no. 1, pp. 93–114, Jan. 2021, doi: 10.1080/13527266.2019.1630664.

[13] M. Smith, *Targeted: How Technology Is Revolutionizing Advertising and the Way Companies Reach Consumers*, 1st ed. New York, NY, USA: AMACOM, Nov. 2014.

[14] S. Yuan, A. Zainal Abidin, M. Sloan, and J. Wang, "Internet advertising: An interplay among advertisers, online publishers, ad exchanges and Web users," 2012, *arXiv:1206.1754*. [Online]. Available: http://arxiv.org/abs/1206.1754

[15] T. Kawaja. *Display LUMAscape*. Accessed: Sep. 23, 2015. [Online]. Available: http://www.lumapartners.com/lumascapes/display-ad-tech-lumascape

[16] V. Toubiana. (2007). *SquiggleSR*. [Online]. Available: http://www.squigglesr.com

[17] B. Liu, A. Sheth, U. Weinsberg, J. Chandrashekar, and R. Govindan, "Adreveal: Improving transparency into online targeted advertising," in *Proc. Hot Topics Netw.*, 2013, pp. 12:1–12:7.

[18] J. Yan, N. Liu, G. Wang, W. Zhang, Y. Jiang, and Z. Chen, "How much can behavioral targeting help online advertising?" in *Proc. 18th Int. Conf. World Wide Web (WWW)*, 2009, pp. 261–270.

[19] M. Aly, A. Hatch, V. Josifovski, and V. K. Narayanan, "Web-scale user modeling for targeting," in *Proc. 21st Int. Conf. Companion World Wide Web (WWW) Companion*, 2012, pp. 3–12.

[20] M. M. Tsang, S.-C. Ho, and T.-P. Liang, "Consumer attitudes toward mobile advertising: An empirical study," *Int. J. Electron. Commerce*, vol. 8, no. 3, pp. 65–78, Apr. 2004.

[21] J. Parra-Arnau, "Pay-per-tracking: A collaborative masking model for Web browsing," *Inf. Sci.*, vols. 385–386, pp. 96–124, Apr. 2017.

[22] J. Parra-Arnau, "Optimized, direct sale of privacy in personal data marketplaces," *Inf. Sci.*, vol. 424, pp. 354–384, Jan. 2018.

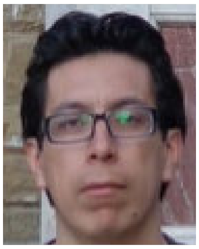
[23] Amazon. *Alexa Top Sites*. [Online]. Available: https://aws.amazon.com/marketplace/pp/Amazon-Web-Services-Alexa-Top-Sites/B07QK2XWNV

[24] T. Micro. *Site Safety Center*. [Online]. Available: https://global.sitesafety.trendmicro.com

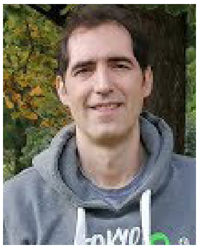
[25] J. P. Achara, J. Parra-Arnau, and C. Castelluccia, "MyTrackingChoices: Pacifying the ad-block war by enforcing user privacy preferences," in *Proc. Annu. Workshop Econ. Inform. Secur. (WEIS)*, Jul. 2016, pp. 1–13, Paper 48.

[26] (Mar. 2016). *Easylist–Overview*. [Online]. Available: https://easylist.github.io

[27] J. Sørensen and S. Kosta, "Before and after GDPR: The changes in third party presence at public and private European Websites," in *Proc. World Wide Web Conf. (WWW)*, 2019, pp. 1590–1600.



JOSÉ ESTRADA-JIMÉNEZ received the bachelor's degree from Escuela Politécnica Nacional (EPN), Quito, Ecuador, in 2007, and the M.S. and Ph.D. degrees from Universitat Politécnica de Catalunya (UPC), Barcelona, Spain, in 2013 and 2020, respectively. His current research interests include encompass data privacy and information security.



JAVIER PARRA-ARNAU received the M.S. degree in telecommunications engineering and the M.S. and Ph.D. degrees in telematics engineering from Universitat Politécnica de Catalunya (UPC), in 2004, 2009, and 2013, respectively. As a Postdoctoral Researcher, he has worked at INRIA, Karlsruhe Institute of Technology, NEC Labs Europe, and Universitat Rovira i Virgili. He is currently a Senior Researcher at UPC. Among other honors, he received the Best Ph.D. Thesis Prize on information and communication technologies in banking from the Official College of Telecommunication Engineers and Banco Sabadell. He was awarded the Postdoctoral Fellowships Alexander von Humboldt, Juan de la Cierva–Formació, and Juan de la Cierva–Inc. He received the prize Data Protection by Design by the Catalan Data Protection Authority, in 2016, and was awarded the second prize Research on Data Protection Emilio Aced by the Spanish Data Protection Agency, in 2018.

ANA RODRÍGUEZ-HOYOS received the bachelor's degree from Escuela Politécnica Nacional (EPN), Quito, Ecuador, in 2010, and the M.S. and Ph.D. degrees from Universitat Politécnica de Catalunya (UPC), Barcelona, Spain, in 2013 and 2020, respectively. Her current research interests include encompass data privacy and machine learning.

JORDI FORNÉ received the M.S. and Ph.D. degrees in telecommunications engineering from Universitat Politécnica de Catalunya (UPC), Barcelona, Spain, in 1992 and 1997, respectively. From 2007 to 2012, he was a Coordinator of the Ph.D. Program in telematics engineering and the Director of the Master's Research Program in telematics engineering. Since 2014, he has been in possession of the Advanced-Research and a Full Professor accreditations. He is currently a Full Professor in telecommunications engineering with the School of Barcelona, UPC, and the Head of the Data Privacy Team, Department of Network Engineering. His research interests include information security and privacy.

ESTEVE PALLARÉS-SEGARRA received the M.S. degree in telecommunications engineering and the Ph.D. degree from Universitat Politécnica de Catalunya (UPC), Spain, in 1994 and 2001, respectively. He is a member of the Smart Services for Information Systems and Communication Networks (SISCOM) Research Group, Department of Networking Engineering, UPC. He is currently an Associate Professor at the Barcelona School of Telecommunications Engineering, UPC. His research interests include span information security and data privacy.