Understanding web pornography usage from traffic analysis

Andrea Morichetta [a,b], Martino Trevisan [a,*], Luca Vassio [a], Julia Krickl [c,d]

[a] Politecnico di Torino, Italy
[b] TU Wien, Austria
[c] University of Vienna, Austria
[d] Austrian Institute for Applied Telecommunications (ÖIAT), Austria

ARTICLE INFO

Keywords:
Web pornography
Network measurements
User behavior

ABSTRACT

Pornography is massively available on the Internet, often free of charge. It represents a significant fraction of the overall Internet traffic, with thousands of websites and millions of users. Studying web pornography consumption is useful to understand human behavior, and it is crucial for different disciplines, helping in sociological, statistical and behavioral research. However, given the lack of public datasets, most of the works build on surveys, limited by multiple factors, e.g., unreliable answers that volunteers may (even unconsciously) give. In this work, we analyze anonymized accesses to pornography websites using HTTP-level traces collected from an operational network. Our dataset includes anonymized traffic from about 15 000 broadband subscribers over three years. We use it to provide quantitative figures on pornographic website consumption, focusing on time and frequency of use, habits, and trends. We also compare web pornography users' interests with those who do not consume web pornography, showing notable differences.

1. Introduction

Pornography and technology have enjoyed a close relationship in recent decades, with multimedia technologies hugely increasing the porn sector's audience. Major technological revolutions led to new forms of expression and access to pornography. From the limited market reachable through public theaters, the introduction of videotapes in the 1970s abruptly changed the way of consuming pornography, allowing access to the content in the privacy and comfort of each person's home. Later, the birth of cable TV and specialty channels in the 1990s allowed a further step toward accessibility and privacy, giving the possibility to retrieve content directly from home, removing the need to pick up the products from a physical store. Finally, the Internet revolutionized the market again, guaranteeing direct delivery to every person with a broadband connection. After a first phase, when the users exchanged files through Peer-to-Peer software [1,2], nowadays, they can interact through forums and webcams, enjoy content free of charge, and, at the same time, anonymity while expanding the audience [3]. Price et al. [4], in a study published in 2016, testify the increase of pornography consumption in the US since 1973. In 2017, the most popular pornographic platform worldwide (Pornhub, according to Alexa ranking) claimed 80 million daily accesses to its website.[1] Thus, it is of no surprise that Internet pornography's role as a prevalent component of popular culture and the importance of its study has been recognized for a long time [5,6].

Many studies focused on web pornography (WP) aim to describe consumption patterns or pinpoint eventual pathologies correlated to excessive use. However, such works typically come from the medical and psychological communities and are based on surveys covering a rather small number of volunteers. Grubbs et al. [7] provide a recent and comprehensive literature review on the subject. Moreover, previous studies [8,9] report that people tend to lie, either consciously or unconsciously, when answering to private-life concerning surveys, especially about sexuality. Indeed, some people declare more WP usage than real (e.g., to show to be uninhibited), while others understate their actual consumption, fearing social blame. These behaviors take the name of social desirability biases and egosyntonic/egodystonic feelings (i.e., being or not in accordance with the self-image). Both of them make surveys less reliable than other sources of information.

In contrast to previous works, in this study, we investigate WP through passive network measurements collected from an Italian Internet service provider's operational network. Our dataset includes anonymized Internet accesses of about 15 000 broadband subscribers over three years. Useful to our analysis, MindGeek, the largest WP

* Corresponding author.
E-mail addresses: andrea.morichetta@polito.it, andrea.morichetta@tuwien.ac.at (A. Morichetta), martino.trevisan@polito.it (M. Trevisan), luca.vassio@polito.it (L. Vassio), a11734102@unet.univie.ac.at, krickl@oiat.at (J. Krickl).

[1] www.pornhub.com/insights/2017-year-in-review.

company operating many popular websites, switched to encryption only in April 2017, becoming the first big player in the WP industry to adopt HTTPS [10]. As such, the vast majority of WP websites used plain-text HTTP up to March 2017, allowing us to leverage HTTP-level measurements and obtain detailed figures of WP consumption. Using recent advances in data science and machine learning, we extract only user actions to WP portals from a deluge of HTTP data, thus discarding uninteresting downloads, e.g., images, scripts, etc. We use these data to understand WP consumption of broadband subscribers from a variety of viewpoints.

In this work, we do not propose novel methodologies to identify Internet pornography. Instead, we refer to the term WP to any online material that, directly or indirectly, *seeks to bring about sexual stimulation* [11]. Therefore, we use the term *pornographic website* to describe services that provide actual pornographic videos, sell sex-related merchandise, help in arranging sexual encounters, etc. We refer only to adult pornography websites, and we do not advocate the inclusion of child pornography websites in our research. Thus, this paper has no application whatsoever to child pornography. The word pornography, in this article context, refers exclusively to content that is legally accessible in the territories of the EU and the US.

Our analysis computes statistics inspired by surveys detailed in sociological and behavioral literature and from WP portal reports. We restrict our analysis only to those that, given our data, we were able to compute. Our results enhance the visibility and the understanding of topics related to WP consumption, giving a less mediated overview of users' behaviors, mostly confirming what emerges from sociological and behavioral surveys by proving a comprehensive representation of WP-related conducts. The main contributions of this paper are:

- Providing a thorough characterization of WP consumption using passive measurements from 15 000 broadband subscribers.
- Showing how users moved to mobile devices through the years, even if the time spent on WP remains constant.
- Showing that typical WP sessions last less than 15 min, with users rarely accessing more than one website. WP sessions are usually longer than sessions on generic websites.

The main findings show:

- Less than 10% of users consume WP more than 15 days on a month, and repeated use within a single day is sporadic. There is not a linear correlation between the time spent on the web and the fruition of WP.
- Users who consume WP are more interested in gaming and technologies than those who do not. WP users present bias toward addictive activities such as online gambling.
- A few WP websites and corporations rule the market, and search engines are the main means to reach WP.

This paper extends our preliminary work [12] in several directions. We compare the characteristics of web browsing for WP and non-WP sessions, showing notable differences. We provide a thorough comparison of WP and non-WP users' behavior in terms of interests, deepening the correlations with possibly risky behaviors, e.g., online gambling. Furthermore, we offer a discussion about our work's implications both from a computer network and a sociological perspective. The remainder of the paper is organized as follows: Section 2 summarizes related work. Section 3 describes data collection, methodology and privacy issues, while Section 4 presents all our results. Finally, Section 5 discusses the implications and limitations of our findings as well as future work, and Section 6 concludes the paper.

2. Related work

Pornography, and, in modern times, web pornography is still seen as a risk and thus to be filtered out. The reason, according to Paasonen [27], can be found in the historical background of censorship and regulation, rooted in Western traditions' moral judgment. Increasingly, WP is investigated as an enabler and mediator of sexual relations in society, manifesting societal relations of gender, sexuality, and power [5]. In feminist research, WP is a topic at the core of the theoretical divide between anti-pornography and porn-sympathetic or pro-sex scholars [28]. Despite the stigma, researchers started addressing the topic from different perspectives, studying the relationship between humans and web pornography and how it is perceived.

Sociology. Most previous works on online pornography study the interaction between users and WP by leveraging the information included in surveys proposed to groups of volunteers. Vaillancourt-Morel et al. [16] examine the potential presence of different profiles of pornography users and their relation to sexual satisfaction and sexual dysfunction. The authors evaluate a poll involving 830 adults and group users' behaviors in three clusters according to the usage of web porn: recreational, highly distressed, and compulsive, each category associated with different reactions. Daspe et al. [13] investigate the relationship between the frequency of pornography consumption and the personal perception of this behavior, pointing out that often there are sizeable discrepancies. Another analysis of the phenomenon is provided by Grubbs et al. [14], who analyze two participant sets, students and adults. They show that moral scruples can infect the self-impression over their consumption. Short et al. [15] propose a critical analysis of WP, showing the various limitations of state-of-the-art studies that estimated WP consumption, concerning its definition, usage, and the uncertainty of its measurements.

Computer science. Other works study WP using network measurements rather than relying on surveys. In 2004, the authors of [6] highlighted the importance of pornography as a principal component of popular culture and the importance of studying, acquiring, and cataloging pornographic websites as part of public library collections. Ortiz et al. [21] study Chilean websites containing human images and classify them in "normal", "porn", and "nude", to automatically discover WP websites. Tyson et al. [24] extract trends and characteristics in a notable adult video portal (YouPorn) by analyzing almost 200 k videos, together with metadata such as page content, ratings, and tags. In a similar direction, Mazieres et al. [20] produce and analyze a semantic network of WP categories extracted from the xHamster portal, finding predominant classes, and inspecting their meaning. Schuhmacher et al. [22] address a similar problem, from a different viewpoint. In their work, the authors inspect nicknames of users commenting on videos of the YouPorn website. They divide nickname classes: male-given name, female-given name, and explicit content, and look for their relation with category tags. This analysis results in an exploration of users' interests to provide suitable recommendations. Coletto et al. [17] study users' activities in social networks related to WP. The goal is to explore these communities and extract information. The authors analyzed the communities' seclusion features and the users' characteristics in terms of age, habits, and gender. Recently, Yu et al. [29] compare in their article WP video consumption to general video streaming services such as YouTube. They find notable differences in terms of video duration and frequency at which users rate videos. Moreover, the views for WP converge around the most popular videos in contrast to videos on YouTube. Zhang et al. [26] analyze the traffic from a vast IPv6 only academic Chinese network, looking for adult content. They filter raw packets with a Naïve Bayes approach, crawling webpages, and analyzing the content. The results show that only a few platforms moved to IPv6. Farelly et al. [18] analyze a broad set of video from the WP website xHamster, and find that they tend to have longer videos compared to mainstream streaming sites and garner more views. xHamster has been studied also by Wong et al. [25] and Song et al. [23]. They use active measurements to collect metadata on almost 4 million unique videos that span the lifetime of xHamster from 2007 to 2018, finding significant differences between adult streaming services and traditional streaming ones. In

Table 1
Related works summary on WP.

Paper	User data	Survey based	Multi websites	Active meas.	Passive meas.	Objective
Sociology						
Daspe et al. [13]	Yes	Yes	–	–	–	WP use and personal perception
Grubbs et al. [14]	Yes	Yes	–	–	–	Moral scruples and self in WP
Short et al. [15]	–	Yes	–	–	–	Review of WP works
V-Morel et al. [16]	Yes	Yes	–	–	–	Categories of WP usage
Computer science						
Coletto et al. [17]	Yes	–	Yes	Yes	–	Explore social network communities
Farrelly et al. [18]	–	–	–	Yes	–	Analysis of WP xHamster
Grammenos et al. [19]	–	–	–	Yes	Yes	Session pattern access
Mazieres et al. [20]	–	–	–	Yes	–	Semantic analysis of xHamster categories
Ortiz et al. [21]	–	–	Yes	Yes	–	Classify chilean pornography images hubs
Schuhmacher et al. [22]	–	–	–	Yes	–	YouPorn categories
Song et al. [23]	–	–	–	Yes	–	WP vs. traditional stream
Tyson et al. [24]	–	–	–	Yes	–	Trends and characteristics in YouPorn
Wong et al. [25]	–	–	–	Yes	–	Analysis of xHamster
Zhang et al. [26]	–	–	Yes	Yes	Yes	Inspect which WP services moved to IPv6
Our work	–	–	**Yes**	–	**Yes**	**Longitudinal analysis of WP use**

Table 2
Dataset description.

Duration	3 years (2014–2017)
Subscribers	$\approx 15\,k$
Log size	20.5 TB
WP websites	59 989
WP visits to webpages	58 238 419
WP sessions	4 135 322
non-WP websites	5 456 846
non-WP visits to webpages	1 176 453 716
non-WP sessions	33 705 684

recent work, Grammenos et al. [19], they gather and analyze data from a significant Content Delivery Network, covering 1 h of access logs for a porn website, combining the study of log data with metadata scraping.

We summarize the related work in Table 1, separating those from the sociological and medical community and those from the computer science field. The table shows how the majority of the latter builds on active network measurements – i.e., the authors automatically download and analyze webpages using the so-called *web crawlers*, and mostly focus on analyzing a single WP portal. The few other works that leveraged passive measurement data, i.e., [17,21,26], focus on circumscribed WP categories or analysis. To the best of our knowledge, we are the first to use passive measurements – i.e., data collected from regular users – to study the consumption of different WP websites over an extended period.

3. Measurements and methodology

3.1. Data collection

In this work, we build on network measurements coming from passive monitoring of a population of broadband subscribers over three years (from March 2014 to March 2017). We have instrumented a Point-of-Presence (PoP) of a European ISP, where the traffic of $\approx$ 10 000 ADSL, and $\approx$ 5 000 FTTH subscribers is aggregated. ADSL downlink capacity is 4–20 Mb/s, with uplink limited to 1 Mb/s. FTTH users enjoy 100 Mb/s downlink and 10 Mb/s uplink. Each subscription refers to an installation, where all users' devices (PCs, smartphones, etc.) connect via WiFi or Ethernet cable through a home gateway. Relevant to our analysis, the ISP provides each subscriber with a fixed IP address, allowing us to track them over time. Nonetheless, a small fraction of subscribers abandoned the ISP during the observation period, and few new ones joined. All ADSL subscribers are residential customers (i.e., households), while a small number of business subscribers exist among the FTTH customers. We report details of our dataset in Table 2. Unfortunately, during the period considered in the paper, the probe suffered some outages, lasting from a few hours up to some months, due to software bugs introduced by updates and hardware failures. As such, the results we present have missing data for those periods, as noticeable in Fig. 1.

To gather measurements, we rely on Tstat [30], a passive meter that builds rich per-flow summaries with hundreds of statistics for each TCP and UDP flow issued by clients. Besides, Tstat integrates a DPI module that creates log files containing details for all the observed HTTP transactions. For each transaction, Tstat records the URL, the client identifier, and other HTTP headers of requests and responses. Our measurements are based on the inspection of HTTP headers and, thus, neglect all encrypted traffic. However, no primary WP portal used encryption at the time we collected the dataset. We copy the generated log files to our back-end servers with a daily frequency. We store data on a medium-sized Hadoop cluster to allow scalable processing. All processing is performed using Apache Spark and Python. The stored data covers three years of measurements, totaling 20.5 TB of compressed and anonymized log files (related to around 138 billion flow records).

3.2. Definition of user and dataset limitations

Our PoP is located at the Broadband Remote Access Server (BRAS) level. A unique and fixed IP address identifies each subscription. However, subscriptions refer to households where potentially more than one person surfs the Internet, sharing the same public IP address. As such, relying on the client IP to identify a user would not be precise enough to study habits and behaviors. Similarly to [31], in our work, we define a *user* as the concatenation of the client IP address and the user-agent as extracted from the corresponding HTTP header. The user agent string is sent in the form of an HTTP request header. It contains information about the agent originating the request, details on the system, e.g., the operating system (and their versions), the client browser, and the content display framework. The use of the user agent allows us to perform analyses in a per-browser fashion – i.e., each user-agent string observed in a household. With this definition, a single person may appear multiple times with different identifiers if they use various devices, or their device incurs software updates that modify the user agent string. Simultaneously, two users with the same machine, running the same operative system, and using the same browser, would be considered the same. Privacy requirements limit any finer granularity.

The evaluated dataset includes only a regional sample of households in a single country. Users in other regions may have diverse browsing habits. Equally, mobile devices have only been monitored while connected to home WiFi networks. As such, our quantification of browsing on mobile terminals is a lower-bound since we do not capture visits performed under mobile networks.

3.3. User action extraction

Starting from an HTTP-level trace, we filter the data to identify those HTTP requests containing an explicit user action. This step aims to isolate users' behavior discarding all HTTP traffic related to internal objects of webpages such as images, style-sheets, and scripts, helping the analyses focus only on the *intentionally* visited webpages. To this end, we rely on the methodology described in our previous work [32] that develops a machine-learning model to pinpoint intentionally visited URLs (or *click stream*) from raw HTTP traces. The employed approach has a core module based on a supervised classifier that can recognize user actions in HTTP traces. It achieves an accuracy of ≈ 98% and can be successfully applied to different scenarios, including smartphone apps [33]. We refer the reader to [32] for a detailed description of our methodology for user action extraction and a comparison with other literature proposals (e.g., [34,35]).

After this phase, we obtain 1.1 billion user actions/visited webpages towards more than 5 million different domains. For each user, we determine information about the operating system in use (e.g., Windows 10), the browser (e.g., Chrome), and if the device was a PC, a smartphone, or a tablet. We extract this information from the original HTTP request user-agent, using the Universal Device Detection library.[2] In the remainder of the paper, we only consider user actions, to which we refer to the term *visited webpages*.

3.4. Session definition and WP filtering

After the extractions of the *intentionally* visited webpages, we perform a further step to identify *sessions* of continuous activity. To this end, we group data by user and process HTTP transactions by time. We then identify a session as follows: when a user accesses a website, we open a new session and account all subsequently visited webpages. We terminate a session if we do not observe any user action for a period of 30 min. While defining a browsing session is complicated [36], we consider a time larger than 30 min as an indication of the end of the session as it is often seen in previous works (e.g., [37]), and in applications like Google Analytics.[3]

Finally, we want to filter only those entries referring to WP websites. Studying innovative methodologies to isolate traffic towards a particular class of services automatically is out of this work scope, and we employ a list-based approach to perform classification. We build on publicly available lists, achieving robustness by combining three different sources.[4] These three lists provide a set of domain names that offer different WP content (ranging from video streaming to thematic forums). To avoid false positives, we consider only those domain names contained in at least two over three lists. We come up with 59 989 unique entries, arranged over 460 top-level domains. We observe an average of 13 k active WP users per month.

Like the general definition of *session* expressed above, we can then define *WP sessions*: we open a new WP session when a user accesses a pornographic website and accounts for it all subsequent WP visited webpages until we do not observe any WP usage for 30 min. In total we obtain more than 4 million WP session in our dataset (see Table 2).

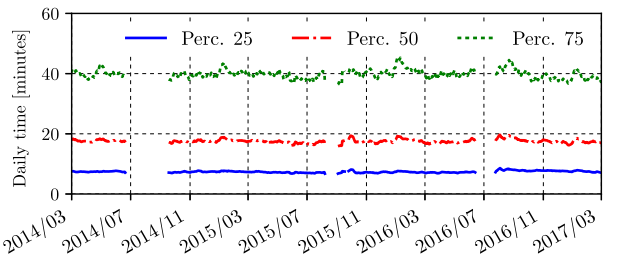
3.5. Privacy and ethical concerns

Passive measurements potentially expose information that may threaten users' privacy [38]. As such, our data collection program has been approved by the partner ISP and by our University's ethical board. Moreover, this specific data analysis project was also subject to a privacy impact assessment.

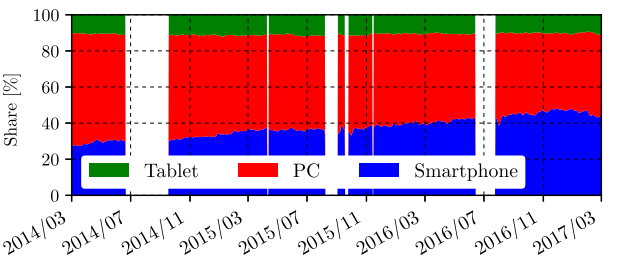
[2] github.com/piwik/device-detector.

[3] support.google.com/analytics/answer/2731565?.

[4] www.shallalist.de/categories.html, www.similarweb.com, and dsi.ut-capitole.fr/blacklists/index_en.php.



(a) Per-user daily time spent on WP. 25^{th}, 50^{th}, and 75^{th} percentiles are shown.



(b) Device category share for accessing WP.

Fig. 1. Usage trends from March 2014 to March 2017.

We undertake several countermeasures to avoid recording any personally identifiable information. Before any storage, all client identifiers (i.e., IP addresses) are anonymized using the Crypto-PAn algorithm [39], and URLs are truncated to avoid recording URL-encoded parameters. We vary encryption keys monthly to prevent persistent user tracking that may leak excessive sensitive information. Private data such as cookies and `Post` data are not monitored at all. We store logs in a secured data center in an encrypted format. We emphasize again that, in our research, we only consider adult pornography websites obtained through open datasets, referring exclusively to content that is legal in the territories of the EU and the USA.

4. Results

In this section, we report the most significant figures that emerge from our analysis. We first focus on the temporal dimension, showing the evolution of WP consumption from 2014 to 2017 in terms of volume and device type. We then characterize WP sessions in terms of duration and frequency and quantify WP pervasiveness in the monitored population. We also show the peculiarities of WP browsing and compare the interests of WP and non-WP users. Finally, we provide some figures about the popularity of services.

4.1. WP consumption trends over the years

Our first analysis describes WP consumption trends from 2014 to 2017. We include notable results in Fig. 1. In Fig. 1a, we show the time spent on WP by monitored users. The blue (solid), red (dash-dot), and green (dashed) curves report, respectively, the 25th, 50th and 75th percentiles of the total per-user daily time spent on WP, i.e., the sum of the duration of all WP sessions. Curves are calculated only for active users, i.e., users visiting at least one WP website in a day. Curves are not continuous due to the lack of data caused by outages in our measurement infrastructure. The outcome shows a rather stable trend over the observation period, with half of the users spending less than 18 min per day on WP; however, almost 25% of users reach 40 min of daily activity. The statistics above have a daily-based outlook on WP activities, and they do not provide figures about the repeated use of

WP across multiple days by the same user, a topic that we will analyze later. Measuring the *fraction* of users accessing WP is not easy using our data, as a single identifier – the client IP address – identifies a broadband subscription, potentially shared by multiple users. However, we notice that every day 12% of subscribers access WP websites, and this value is constant across the years. We provide further analysis of WP pervasiveness in Section 4.4. As it emerges from our previous work [2], during the same period, the daily traffic per broadband customer has increased at a constant rate, almost doubling from 2014 to 2017. We do not see a similar increase in WP traffic.

We can compare these results with statistics from surveys, fortifying or confuting what the participants declare. For instance, Vaillancourt-Morel et al. [16] study the characteristics of WP users, showing that the majority of the chosen sample uses WP for recreation only, on average for 24 min per week. On our dataset, we find that, on average, a user accessing WP spend 37 min per week. Despite the different user bases, this result can be seen in the user's tendency to lie when answering surveys [8,9].

Then, we investigate the evolution in device category use (PCs, tablets, and smartphones). We compute, for each device category, its share in terms of the number of sessions. Fig. 1b shows the results. We notice that though PCs and laptops (red surface) still cover a relevant fraction of visits in 2017, smartphones (blue surface) have largely increased their share from 27% to 42% at the expense of PCs. The volume of tablets, reported in green, is instead rather constant. Not shown for brevity, the evolution of the daily time spent using different devices did not change consistently throughout the years (see Section 4.2 for more details). Not visible in the picture, the absolute number of users with PCs remained more or less constant throughout the years, while smartphone users increased by about 32%. Finally, the other metrics we will explore in the next sections did not exhibit significant trends worth mentioning here. As there is no significant increase in the usage over the years, interesting conclusions could be drawn in terms of the "addictiveness" of pornography itself. Constant availability on the smartphone does not significantly increase the WP consumption in the sample over the observed period of time.

Take away: *the overall WP consumption remained constant over the years, and stands, in median value, on 18 min per day. Still, we see a sharp increase in smartphone usage.*

4.2. Characterizing WP sessions

As previously shown, WP consumption for each user is rather stable across the years. As such, we now restrict our investigation to one month, allowing more straightforward data processing without sacrificing the analysis accuracy. Therefore, in the remainder of the paper, we restrict the study to the last month of our dataset that neither includes public holidays nor measurement outages, i.e., October 2016.

In this section, we characterize WP sessions, separately for the three device categories, showing the results in Fig. 2. We first characterize WP sessions in terms of duration — given the definition of a session as continuous access to WP, given a 30 min timeout. Fig. 2a shows the empirical cumulative distribution function (CDF) of session duration, expressed in minutes. The duration is larger for PCs than for tablets and smartphones. While most of the sessions are rather short, i.e., less than 15 min for PCs and 10 for smartphones, we observe sporadic longer sessions up to one hour or more, with the 75th percentile reaching 22 min. On a global scale, Pornhub has found similar results.[5] The average session time reported by Pornhub for Italy is 9 min and 30 s, similar to what we observe from our analysis.

We now draw the attention to the number of webpages accessed within WP sessions, whose CDF is reported in Fig. 1b. Remind that

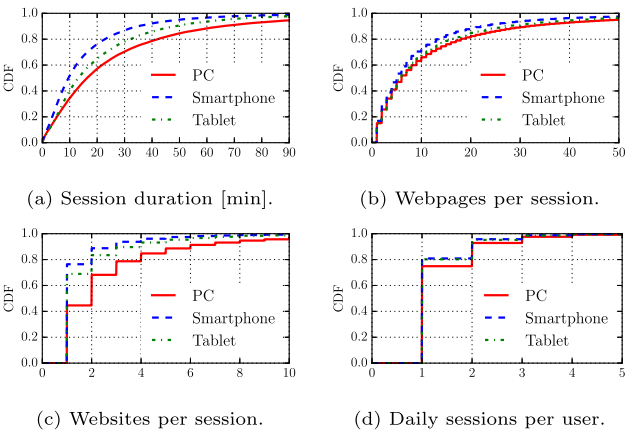


Fig. 2. CDF of WP session characteristics, divided by device type.

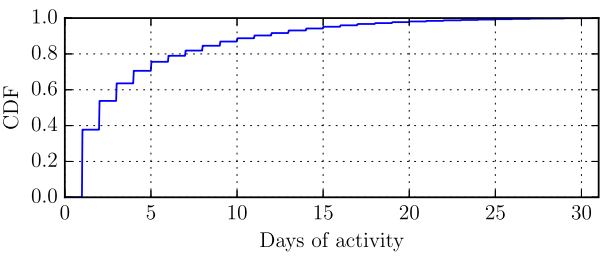


Fig. 3. Number of distinct days in which users consumed WP in a month.

this number is calculated only for explicit user actions, thus discarding other HTTP transactions for images, scripts, and others. Here the difference among devices is limited, with users accessing in median 5 or 6 webpages in a session, with 28% of them limited to one or two. However, in 30% of the sessions, more than 10 webpages are visited in one session. Similarly, in Fig. 2c, we report the distribution of the number of *unique* websites accessed during a WP session.[6] Results show that smartphone users tend to focus on a single WP website at a time (78% of sessions). In contrast, PC users are more prone to visit multiple websites. For all the devices, very few sessions include visits to 4 or more different websites.

Finally, Fig. 2d reports the number of daily sessions for an active user. Given a user that accesses WP content on a particular day, we count the number of sessions they undertake. The figure shows that users hardly make repeated use of WP within a day, without differences among devices. Only in 20% of cases, we observe repeated use, with a higher number of daily sessions related to more marginal behaviors. Please acknowledge that we start counting from one, as our data hardly allows us to estimate the number of monitored *users* not accessing WP content.

Take away: *During WP sessions, users tend to remain on a single website, for a median time of 10–15 min, depending on the adopted device. A user that accesses WP, rarely does so more than once per day.*

4.3. Usage frequency and seasonality

We now focus on the frequency of WP consumption over the considered month. In Fig. 3, we report the CDF of the number of days of activity of the WP users in the dataset. The figure indicates that the monthly frequency is generally low, with 76% of the users visiting WP five or fewer days in a month. Still, some users show reiterate usage, 8% of them consuming WP more than 15 times a month. These results

[5] See footnote 1.

[6] We easily identify websites using the *hostname* extracted from URLs.

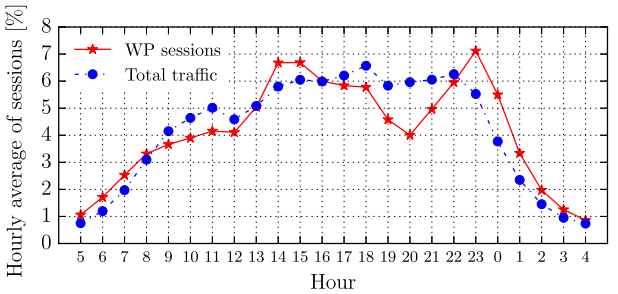


Fig. 4. Average hourly percentage of number of WP sessions and total traffic.

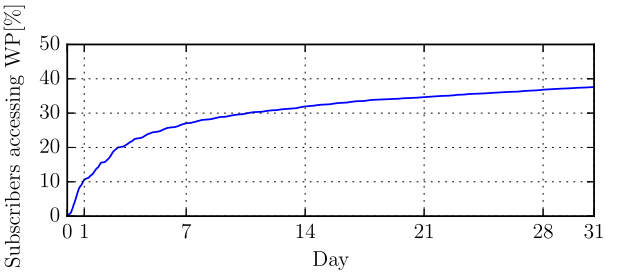


Fig. 6. Cumulative percentage of subscriptions accessing WP at different time in the trace.

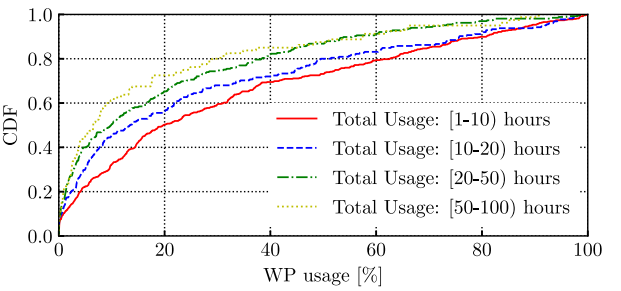


Fig. 7. Share of time spent on WP by users, grouped in different usage classes.

confirm what is found by Daspe et al. [13], who show that the 73% of the participants to a survey access pornography no more than once or twice per week, and only 11% more than five times per week. Given the nature of our dataset, we cannot estimate the number of users *not* consuming WP. Still, an analysis of per-*subscription* traffic to WP is provided in Section 4.4.

The volume of WP sessions also varies during the hours of the day. Fig. 4 provides the average percentage of sessions across the 24 h of the day (red solid line). For ease of visualization, we start the x-axis from 4 A.M., corresponding to the lowest value of the day. The two higher peaks are immediately after lunchtime (2 P.M.–4 P.M.) and after dinner (9 P.M.–midnight). In addition to WP traffic, the figure also reports the overall trend considering all WP and non-WP HTTP transactions, regardless of their nature (dashed blue line). We notice some discrepancies comparing WP to the total traffic; the peaks do not overlap, and the latter is more balanced over daylight hours. A hypothesis for those differences may be related to the fact that accessing pornographic websites is likely to be a private and leisure activity confined to intimate moments, linked to intimacy, and restricted by social rules. We also provide a breakdown across both hours and days of the week, with Fig. 5 showing the heatmap of the percentage variations from the gross weekly average (white color). Warmer tones register values below average, while colder ones show values above. Notice some clear diminishing traffic on Saturday evening (7 P.M.–midnight) and some increased traffic on Saturday, Sunday, and Monday morning (9 A.M.–1 P.M.). Indeed, many commercial activities are closed on Monday morning in the monitored country, perhaps influencing this behavior. Again, Pornhub data shows comparable results, with their heatmap having peaks of traffic in more or less the same time frames (2 P.M.–5 P.M.) and (10 P.M.–midnight). Considering the cumulative daily access, Mondays register the highest values and Saturdays the lowest.

Take away: *Overall, 76% of the users visit WP five or fewer days in a month. Consumption has two peaks after lunch and after dinner. It is steady during working days, while we observe a decrease on Saturday night and an increment on Saturday, Sunday, and Monday morning.*

4.4. WP pervasiveness

We now provide a general analysis of the *fraction* of monitored subscribers consuming WP websites. Unfortunately, our dataset does not contain fine-grained information about WP pervasiveness, being the client IP address shared by all users surfing the web from a subscription, i.e., a household. Still, we can show the fraction of subscriptions where at least one user accessed WP during our observation period. In Fig. 6, the x-axis represents the 31 days of our reference month (being day 1 October 1st, 2016 and day 31 October 31st, 2016), while y-axis reports the cumulative fraction of subscriptions that accessed at least one WP website. Considering the first day, less than 12% of

subscriptions accessed WP, but this fraction raises to 27% after a week. At the end of the month, it reaches 38%, meaning that more than one subscription over three generated traffic towards WP websites at least once in a month. For comparison, 45% and 3% of subscribers

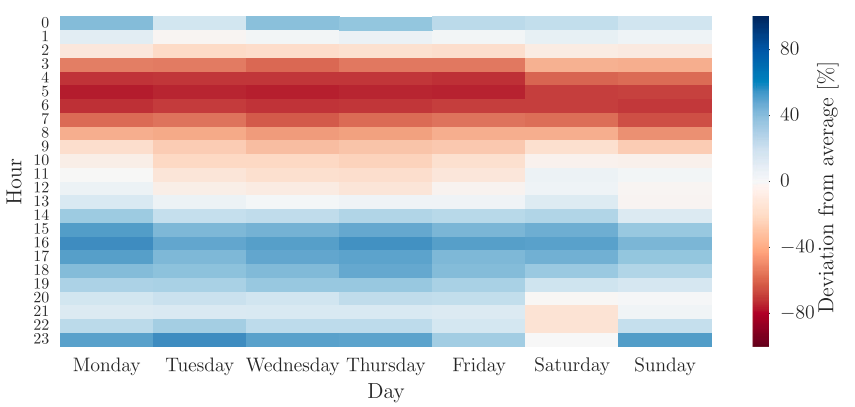
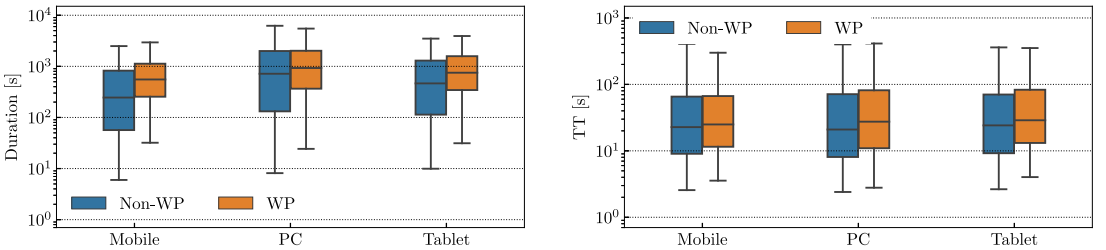


Fig. 5. Weekly breakdown of hourly WP usage. Heat-map of deviation from hourly average.



(a) Session duration. (b) Thinking Time (TT).

Fig. 8. Comparison of session duration and thinking time related to WP and non-WP fruition.

access YouTube and Netflix daily, respectively. Considering social networks, 60% and 25% of subscribers contact Facebook and Instagram daily, respectively.[7] Finally, we observe that this picture did not evolve from a temporal perspective. As we mentioned in Section 4, overall, 12% of users are observed accessing WP on a daily basis with basically no variation over the years. Also, the number of users accessing WP weekly and monthly is rather flat over time.

Take away: *Weekly, 27% of households access WP at least once, while 38% do so monthly.*

4.5. WP vs. non-WP browsing

We now explore the peculiarities of WP browsing in contrast to non-WP activity through the comparison of different characteristics.

Firstly, we investigate the pervasiveness of WP browsing in different users' classes, grouped by the monthly amount of time spent on the web. To this end, given a WP user, we sum the duration of all their browsing sessions (WP and non-WP sessions). Then, we group users into four bins for different time spent on the web and compute for each one the fraction of time spent on WP individually. Fig. 7 depicts the portion of time dedicated to WP in four monthly usage bins, namely $[1 - 10)$ h (red solid line), $[10 - -20)$ h (blue dashed line), $[20 - -50)$ (green dot-dashed line), and $[50 - -100)$ h (yellow dotted line). Interestingly, the figure shows a clear decrease in the fraction of time dedicated to WP when the overall amount of time spent on the web increases. For instance, the median time spent on WP from users that browse the web between 10 and 20 h is 15%, while for users that browse for more than 50 h, the median is only 8%. The absolute value of time spent on WP, not shown here for brevity, slowly increases with overall time. Hence, these results highlight how the time spent for WP fruition is growing less than linearly concerning the total browsing time. Bringing together these results with the analysis reported in Fig. 3, we can assume that, for the monitored population, WP browsing is an occasional activity. Except for a small fraction of the population, access to WP resources is quantitatively limited and not growing proportionally to the rest of the traffic.

Subsequently, we consider the characteristics of browsing sessions, aiming at comparing WP with non-WP consumption. We analyze two metrics, namely the session duration and the thinking time (TT), i.e., the period (in seconds) between two webpage visits within a session. Fig. 8 shows separately the distributions for the device classes, i.e., Mobile, PC, and Tablet. We use boxplots in which the boxes span from the 1st to the 3rd quartile, while whiskers report the 5th and the 95th percentiles; black strokes represent the median.[8]

Fig. 8a illustrates the distribution of session duration comparing WP and non-WP.[9] The figure shows that browsing is generally a concise activity, irrespective of the type of content, with session rarely lasting more than half an hour. Considering all traffic, PCs sessions are longer (12 min in median), while those on mobile phones are shorter (only 5 min in median). Tablets stay in the middle. We notice that WP sessions are generally longer than non-WP. Indeed, WP sessions last 10–15 min in the median, depending on the user device, while non-WP in the order of 5–12 min, suggesting that users that choose to devote their time to WP, generally invest more time than for classical browsing. We put this result in the perspective of the WP website popularity, which reveals that the most accessed WP are video portals (see Section 4.7). Indeed, video consumption is intuitively a longer activity than browsing, and WP online videos are on average longer than non-WP ones [18].

Fig. 8b depicts the distribution of TT for the different device classes. Recall that TT is the amount of time between two user actions within a session. It is worth to mention that the non-WP category contains user actions over many different content topics. Thus, what we see in the output is an average of users' behavior while browsing. The figure shows minimal differences between device classes and WP/non-WP. In general, PC sessions have slightly shorter TT than mobile and tablets, with differences in the order of a few seconds, looking at the median. WP has longer median TT for all devices classes, but again few seconds separate the distributions. This result suggests that users have a similar approach for WP and non-WP in terms of time spent on a webpage, i.e., before opening the following page.

Take away: *There is a less than a linear correlation between the time spent on the web and the fruition of WP. WP browsing exhibits peculiar characteristics, with longer sessions. We found negligible differences in terms of thinking time between webpage accesses.*

4.6. Interests of WP and non-WP users

We here focus on users' behavior in terms of personal interests, comparing those who consume WP with those who do not. Non-WP users are those that do not access a single WP content (considering October 2016). We characterize their interests using the list of websites they visit. We categorize each visited website using the *Investigate* API of the Cisco Umbrella platform.[10] This service provides a categorization into a topic based on the content of the website (e.g., News). We consider users interested in a topic if they access at least one website belonging to it. This technique allows us to compute the share of users interested in each class. We consider 11 topics, from generic News and Sport to topics related to specific user behaviors such as Gaming and Gambling.

We show the results in Fig. 9 using a radar plot, separately for WP and non-WP users. The two lines report the percentage of users interested in a specific topic. In general, WP users consume more diverse content: in fact, the WP users' curve covers a larger area than

[7] The reader can find a deeper analysis in our previous work [2].

[8] The choice of using boxplots instead of plotting CDFs has the goal of easing the readability of the results and avoiding overlapping lines.

[9] For the session duration of WP browsing, refer to Fig. 2a.

[10] https://docs.umbrella.com/investigate-api/docs/introduction-to-cisco-investigate.

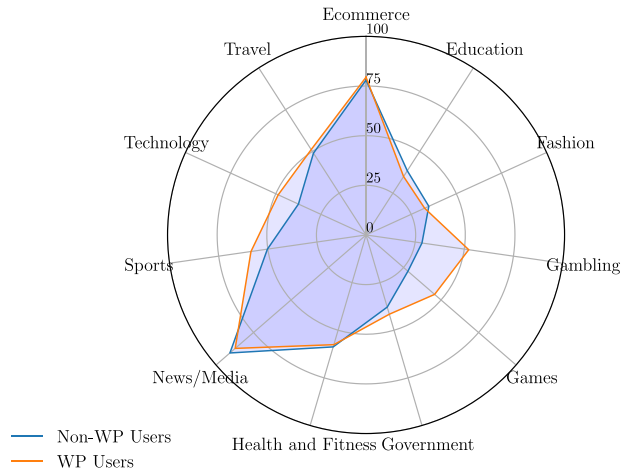


Fig. 9. Interests of WP and non WP users.

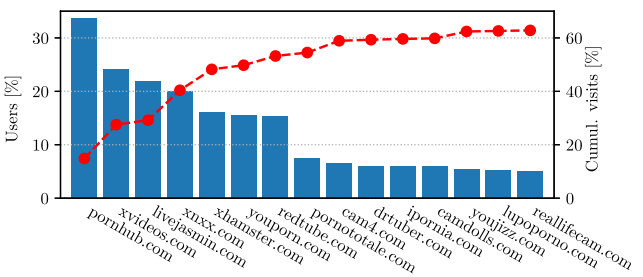


Fig. 10. Top-15 WP websites ranked according to percentage of users accessing them. Cumulative percentages of their visits with respect to all WP visits are also shown.

the non-WP curve. While, for some topics, we do not notice significant differences (News, E-commerce, Health, and Fitness), in some cases, users' interests considerably diverge. Significant differences hold for Gambling, Games, and Technology classes. We observe that 51% of WP users access Gambling websites, while this happens in only 27% of non-WP. This result is particularly interesting in the perspective of other studies that establish a connection between WP and other kinds of addiction, e.g., to gambling or drugs, or association to violent thinking or conducts, e.g., search for weapons [40]. We also investigated other kinds of biases, like the connection to discriminating ideas, such as racism, but the data points were too few to draw any statistically significant conclusion.

Take away: *WP users are more prone to visit websites related to gaming and technology. Moreover, they are considerably more interested in addiction-prone services, like gambling websites.*

4.7. Most contacted WP websites

This section briefly illustrates the most popular WP websites and how the traffic is distributed among them. Similarly to the global Internet trend, a few big players dominate the market. Looking at the Alexa rank for 2016, three WP websites appear among the top-50, namely pornhub.com, xvideos.com, and livejasmin.com, with the first one, ranked 29th, just behind linkedin.com, and the other two respectively 47th and 49th. Considering our dataset, we observe similar positioning for WP websites, with top-tier WP aggregators leading the rank. In Fig. 10, we show the percentage of users reached by the top-15 WP websites using bars (left-most y-axis) and the cumulative percentage of visits to these services (red line, right-most y-axis). We first note that the top-15 websites are all video-based WP services, some of them also offer chatting functionalities, but we do not find any forum representative. In total, users accessed 7 048 different websites during the entire month. The top-3 websites in our dataset match exactly

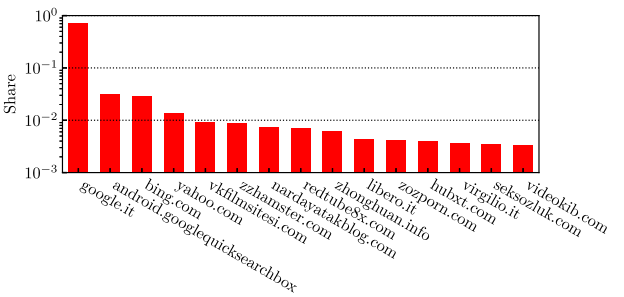


Fig. 11. Top-15 websites present in the referer HTTP headers on the beginning of WP sessions. Notice the log scale.

the Alexa ranking, with pornhub.com being accessed by 34% of users. Global tendencies are reflected in our top-15, with only two websites being local representatives of the monitored country (pornototale.com and lupoporno.com). The red line reporting the cumulative percentage of WP visits supports us to highlight significant results. The leading WP webpage pornhub.com accounts alone for 14% of total accesses, while the top-15 together account for approximately 63% of all WP visits. Even if not reported in Fig. 10, the percentage reaches 90% considering the top-204 websites, confirming the concentration of users around top services. Interestingly, very similar numbers hold true for the overall traffic observed in our dataset (including also non-WP websites), with the top-15 accounting for 61% of traffic and 90% due to 195 websites.

We are also interested in studying the behavior of the WP website popularity distribution to check if it follows a power law. A power-law probability density function is defined as $p(x) = c \cdot x^{-\alpha}$. We consider all WP visits in the dataset to check this behavior, hence more than 58 million visits towards almost 60 thousand WP websites. Accurately fitting a power-law distribution to empirical data, as well as measuring the goodness of that fit, is not trivial [41]. We are interested in studying the tail of the distribution (i.e., the few very popular websites), and we check if and from what minimal value x_{min} the scaling relationship of the power law begins. We use the methods described in [41] to find the optimal value of x_{min} by creating a power-law fit starting from each unique value in the dataset, then selecting the one that results in the minimal Kolmogorov–Smirnov distance between the data and the fit. In our case, we obtained a value of x_{min} equal to 1 501, and a high significance value, meaning that the distribution follows a power law after that value. This indicates that the WP website popularity distribution has an "heavy-tail" that follows a power law, similar to other natural phenomena. Later in Section 5, we put this result in the perspective of the studies on the Web ecosystem.

We notice that 3 out of 15 WP websites represented in Fig. 10 belong to MindGeek, a company owning pornhub.com, redtube.com, youporn.com, and dozens of other websites [42]. MindGeek websites account for more than 20% of accesses in our dataset, making it the market leader. For comparison, the following website in terms of users and visits is xvideos.com (owned by WGCZ Holding), with less than half of the users of MindGeek services, according to our data, suggesting again a scenario where the ecosystem is lead by a few big players in a dominant position.

Finally, we investigate how users reach WP websites — i.e., from which search engine or aggregator website. To this end, we employ the `Referer` header of HTTP requests, which reports the URL of the previously accessed website. This mechanism allows a web server to identify where users are coming from, from a page containing a link to the website. For our analysis, we use the `Referer` of the first HTTP request of a WP session. About 29% of these requests are without `Referer`, hence in these cases users request the first WP webpage in a new window/tab of the browser or starting from bookmarks. We then consider the websites that are found in the `Referer` header, and show results in Fig. 11. As expected, most users access WP through the Google search engine (71.8%). In the second position, we find (3.1%) which

means that users search for WP directly on the search bar of Android phones. We then find other popular search engines such as Bing (2.8%) and Yahoo (1.3%). We observe Italian aggregator portals that include a web search service, namely Libero (0.4%) and Virgilio (0.3%). Curiously, the remaining 9 websites appear to be temporary WP portals that were neglected by our list – e.g., zzhamster.com or redtube8x.com. They monetize users that navigate through them to reach the legit WP website. As such, they are prosecuted by legitimate content providers and often blocked by national authorities. This controversy is why they were not included in our WP lists, and none of them can, at the time we write this article, be accessed anymore from Europe or the US.

Take away: *Few websites are responsible for the majority of accesses, with the tail of the number of visits following a power law. WP replicates the overall web scenario, where a few big players lead the ecosystem in a dominant position. Search engines, and in particular Google, are the primary means to reach WP websites.*

5. Implications and limitations

In this section, we discuss the implications of our work from a twofold perspective. We first analyze how our results fit in the context of the web ecosystem and then focus on the implications for the study of human sciences in general. Finally, we discuss the limitations of our work in terms of employed datasets (and generalizability of our results) and give the basis for future work.

Implications for computer science and networking studies. To the best of our knowledge, our work is the first attempt to study the WP ecosystem using passive measurements. Indeed, we study how a population of broadband subscribers consume WP, and, in contrast to other works relying uniquely on active measurements, we can characterize the behavior of the users also while accessing non-WP content. Our results show that WP ecosystem follows the same empirical law of the web in general. Few popular websites attract most of the traffic, while there is a long tail of infrequent ones. The web ecosystem has already been proved to follow a power law distribution [43], also when restricting to online video popularity [44], Wikipedia pages [45] and Peer-to-Peer content [46]. Here, we show that the WP sub-ecosystem follows a similar law as well (see Section 4.7). Our work also confirms the rise of mobile devices already reported in industrial reports [47] and research papers [33]. Indeed, Section 4.1 shows that mobile traffic has largely increased their share from 27% to 42% at the expense of PCs.

Implications for social sciences. The body of research on the impact of pornography on users as well as on society at large is divided across ideological lines and shows significant biases, which currently preclude internally valid causal conclusions on the effects of its use, according to Peter et al. [48]. Hence, the importance of research that provides a playing field for further sociological, psychological as well as psycho-social studies on the detailed effects and concrete use of WP. This paper creates this playing field. The effects, moral and legal implications, of excessive use of WP are still largely debated by various sociological, pedagogical and psychological works today. Noteworthy are the main arguments on whether or not users, as well as performers of pornographic acts, are harmed in the production or consumption of pornographic material [49]. Historically, this debate was influenced by 19th and 20th century anti-obscenity laws, and later mostly held by feminist activists and scholars in Europe and the United States [49]. Recently, these debates shifted online and are embodied in part by the growing ‘‘noFap’’-movement, a group of mostly male WP users that restrict the use of pornographic material and claim to gain strength and control over their lives through restriction of masturbatory practices [50]. Contrastingly, sex-positive activists promote sexual freedom and expression as a way of women’s liberation [49]. Helped by streaming platforms and websites such as OnlyFans.com, sex workers are gaining independence from producers and are the owners of their self-produced content. The debate on the effects of WP on society is far from over, as illustrated here, hence the importance of *unbiased* material which will help to contextualize further WP use and its effects on individual behavior as well as society at large. Sociological studies are largely based on categorization and attaching research to individual behavior. The presented paper offers the opportunity to study the behavior regarding WP use in an ideologically neutral way, examining anonymized traffic in a central European context. Anonymization here can be a means to avoid gender and age biases. Clearly, it can only be a starting point for further research, but, however, we find a compelling one.

Limitations. The paper analyzes the activity of approximately 15 000 subscribers. Despite the not negligible size especially if compared to the classical survey approaches, the considered population is just a fraction in the global set of WP users. Thus, the analysis is limited by the partial view obtainable from the study of our sample. The limitations appear in different dimensions. Given the nature of our data, identified by passive traces, it is impossible to obtain a more fine-grained characterization of the subscriber. This constraint prevents us from safely asserting the distinction of users within a household. Furthermore, compared to classical surveys, we have no capability to select a balanced or *randomized* set of users. This shortcoming limits a thorough sociological analysis of different groups across different demographics, such as gender, age, and education. Finally, our data describe the activity of a group of subscribers in a European country. Different continents and cultures come with different habits, preferences, and motivations for internet content consumption. The lack of data on WP use in other areas of the world limits the spectrum and our study’s generalization.

Future research. The number of *interactive* pornographic streaming portals like OnlyFans.com has increased drastically in the past years. Future research could compare classic pornographic websites with little to no user interaction, in the form of chats and the like, to interactive streaming services in which users can directly communicate with the performers by giving tips or expressing wishes. Studies in this direction could help understand fringe behaviors, like daily WP use, for up to one hour. The steady percentage of regular, daily users could be using private chat rooms and streaming portals, granting them exclusive access to a performer/sex worker. These fringe behaviors could be studied in depth while relying on the presented work.

Sociologically, the map of interests can be studied further to identify risky behaviors and dissect behavioral patterns of excessive WP users. The findings could help prevent and alleviate perceived pornography-addiction and identify which behaviors are problematic for the daily lives of WP users. The link between risky behavior, e.g., online gambling and WP use, should be investigated further; the presented paper offers a starting point for research.

6. Conclusion

In this paper, we offered a quantitative analysis of the WP consumption by 15 000 broadband subscribers. To the best of our knowledge, we are the first to use passive network measurements to study users’ interactions with web pornography services. We followed an exploratory approach on data, focusing on questions, topics, and metrics typically requested and analyzed in previous surveys and research works, e.g., frequency of use, and the time spent on WP.

Our results are useful for researchers studying web service consumption and human behavior at large. Furthermore, the chosen metrics allow a comparison with the outcomes of previously conducted surveys and mostly confirmed their results. In general, we observe that users consume WP in sessions of 10-15 min and rarely do more than once per day. Accesses concentrate on two peaks, i.e., after lunch and dinner. WP browsing sessions are generally longer than other sessions, and users consuming WP have different behaviors with regards to non-WP users, with more interest in gaming, technology, and online gambling. Finally, WP traffic volume converges around a few big players, following the

Internet's general trend. This work contributes to understanding actual monitored user behavior, without having to consider survey biases. It aims at providing data to be used in further research on web pornography use and linked behaviors in various disciplines. Moreover, it provides data outside of the influence of private actors, which might be subjected to bias in the course of protecting customers or company revenues.

CRediT authorship contribution statement

Andrea Morichetta: Project administration, Formal analysis, Software, Validation, Writing - original draft. **Martino Trevisan:** Software, Validation, Writing - original draft. **Luca Vassio:** Data curation, Conceptualization, Writing - review & editing. **Julia Krickl:** Project administration, Writing - review & editing.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

The research leading to these results has been funded by the Vienna Science and Technology Fund (WWTF), Austria through project ICT15-129 (BigDAMA) and the Smart-Data@PoliTO center for Big Data technologies. The data collection and analysis has been made in the premises of the Politecnico di Torino.

References

[1] Naimul Basher, Aniket Mahanti, Anirban Mahanti, Carey Williamson, Martin Arlitt, A comparative analysis of web and peer-to-peer traffic, in: Proceedings of the 17th International Conference on World Wide Web, 2008, pp. 287–296.
[2] M. Trevisan, D. Giordano, I. Drago, M.M. Munafò, M. Mellia, Five years at the edge: Watching internet from the ISP network, IEEE/ACM Trans. Netw. (2020) 1–14.
[3] JoAnn Di Filippo, Pornography on the web, in: Web. Studies: Rezuiring Media Studies for the Digital Age. London: Arnold, 2000, pp. 122–129.
[4] Joseph Price, Rich Patterson, Mark Regnerus, Jacob Walley, How much more XXX is generation X consuming? Evidence of changing attitudes and behaviors related to pornography since 1973, J. Sex Res. 53 (1) (2016) 12–20.
[5] Chris Brickell, Sexuality, power and the sociology of the internet, Curr. Sociol. 60 (1) (2012) 28–44.
[6] Juris Dilevko, Lisa Gottlieb, Selection and cataloging of adult pornography web sites for academic libraries, J. Acad. Librariansh. 30 (1) (2004) 36–50.
[7] Joshua B. Grubbs, Paul J. Wright, Abby L. Braden, Joshua A. Wilt, Shane W. Kraus, Internet pornography use and sexual motivation: A systematic review and integration, Ann. Int. Commun. Assoc. 43 (2) (2019) 117–155.
[8] Richard C. Lewontin, Sex, lies, and social science, N. Y. Rev. Books 42 (7) (1995) 24–29.
[9] Eric P. Ochs, Yitzchak M. Binik, The use of couple data to determine the reliability of self-reported sexual behavior, J. Sex Res. 36 (4) (1999) 374–384.
[10] Brian Fung, Porn websites beef up privacy protections days after congress voted to let ISPs share your Web history, 2017, The Washington Post, https://www.washingtonpost.com/news/the-switch/wp/2017/03/30/porn-websites-beef-up-privacy-protections-days-after-congress-voted-to-let-isps-share-your-web-history, 2017-03-30.
[11] Martha Cornog, Libraries, erotica, pornography, Libr. Q.: Inf. Community Policy 61 (4) (1991) 457–459.
[12] Andrea Morichetta, Martino Trevisan, Luca Vassio, Characterizing web pornography consumption from passive measurements, in: David Choffnes, Marinho Barcellos (Eds.), Passive and Active Measurement, Springer International Publishing, Cham, 2019, pp. 304–316.
[13] Marie-Eve Daspe, Marie-Pier Vaillancourt-Morel, Yvan Lussier, Stephane Sabourin, Anik Ferron, When pornography use feels out of control: The moderation effect of relationship and sexual satisfaction, J. Sex Marital Ther. 44 (4) (2018) 343–353.
[14] Joshua Grubbs, Joshua Wilt, Julie Exline, Kenneth Pargament, Shane Kraus, Moral disapproval and perceived addiction to internet pornography: a longitudinal examination, Addiction 113 (3) (2014) 496–506.
[15] Mary Short, Lora Black, Angela Smith, Chad Wetterneck, Daryl Wells, A review of internet pornography use research: Methodology and content from the past 10 years, Cyberpsychology Behav. Soc. Netw. 15 (1) (2012) 13–23.
[16] Marie-Pier Vaillancourt-Morel, Sarah Blais-Lecours, Chloé Labadie, Sophie Bergeron, Stéphane Sabourin, Natacha Godbout, Profiles of cyberpornography use and sexual well-being in adults, J. Sex. Med. 14 (1) (2017) 78–85.
[17] Mauro Coletto, Luca Maria Aiello, Claudio Lucchese, Fabrizio Silvestri, Adult content consumption in online social networks, Soc. Netw. Anal. Min. 7 (1) (2017) 28:1–28:21.
[18] Benjamin Farrelly, Yiying Sun, Aniket Mahanti, Mingwei Gong, Video workload characteristics of online porn: Perspectives from a major video streaming service, in: 2017 IEEE 42nd Conference on Local Computer Networks (LCN), IEEE, 2017, pp. 518–519.
[19] Andreas Grammenos, Aravindh Raman, Timm Böttger, Zafar Gilani, Gareth Tyson, Dissecting the workload of a major adult video portal, in: International Conference on Passive and Active Network Measurement, Springer, 2020, pp. 267–279.
[20] Antoine Mazières, Mathieu Trachman, Jean-Philippe Cointet, Baptiste Coulmont, Christophe Prieur, Deep tags: toward a quantitative analysis of online pornography, Porn Stud. 1 (2) (2014) 80–95.
[21] Javier Ruiz-del Solar, Victor Castañeda, Rodrigo Verschae, R. Baeza-Yates, Felipe Ortiz, Characterizing objectionable image content (pornography and nude images) of specific web segments: Chile as a case study, in: Third Latin American Web Congress (LA-WEB'2005), IEEE, 2005, p. 10.
[22] Michael Schuhmacher, Cäcilia Zirn, Johanna Völker, Exploring Youporn Categories, Tags, and Nicknames for Pleasant Recommendations, ACM, 2013.
[23] Yo-Der Song, Mingwei Gong, Aniket Mahanti, Measurement and analysis of an adult video streaming service, in: 2019 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM), IEEE, 2019, pp. 489–492.
[24] Gareth Tyson, Yehia Elkhatib, Nishanth Sastry, Steve Uhlig, Measurements and analysis of a major adult video portal, ACM Trans. Multimed. Comput. Commun. Appl. 12 (2) (2016) 35:1–35:25.
[25] Cameron Wong, Yo-Der Song, Aniket Mahanti, YouTube of porn: longitudinal measurement, analysis, and characterization of a large porn streaming service, Soc. Netw. Anal. Min. 10 (1) (2020) 1–19.
[26] Shize Zhang, Hui Zhang, Jiahai Yang, Guanglei Song, Jianping Wu, Measurement and analysis of adult websites in IPv6 networks, in: 2019 20th Asia-Pacific Network Operations and Management Symposium (APNOMS), IEEE, 2019, pp. 1–6.
[27] Susanna Paasonen, Ubiquitous yet filtered: Porn and the search, in: Search and Exploration of X-Rated Information: WSDM'13 Workshop Proceedings: February 5, 2013, Rome, Italy, 2013, pp. 13–14.
[28] Karen Boyle, Feminism and pornography, in: The SAGE Handbook of Feminist Theory, 2014, pp. 215–231.
[29] Raymond Yu, Callan Christophersen, Yo-Der Song, Aniket Mahanti, Comparative analysis of adult video streaming services: Characteristics and workload, in: 2019 Network Traffic Measurement and Analysis Conference (TMA), IEEE, 2019, pp. 49–56.
[30] Martino Trevisan, Alessandro Finamore, Marco Mellia, Maurizio Munafò, Dario Rossi, Traffic analysis with off-the-shelf hardware: Challenges and lessons learned, IEEE Commun. Mag. 55 (3) (2017) 163–169.
[31] Zied Ben Houidi, Giuseppe Scavo, Stefano Traverso, Renata Teixeira, Marco Mellia, Soumen Ganguly, The news we like are not the news we visit: News categories popularity in usage data, in: Proceedings of the International AAAI Conference on Web and Social Media, Vol. 13, 2019, pp. 91–102.
[32] Luca Vassio, Idilio Drago, Marco Mellia, Detecting user actions from HTTP traces: Toward an automatic approach, in: 2016 International Wireless Communications and Mobile Computing Conference (IWCMC), IEEE, 2016, pp. 50–55.
[33] Luca Vassio, Idilio Drago, Marco Mellia, Zied Ben Houidi, Mohamed Lamine Lamali, You, the web, and your device: Longitudinal characterization of browsing habits, ACM Trans. Web 12 (4) (2018) 24:1–24:30.
[34] Aniket Mahanti, Carey Williamson, Niklas Carlsson, Martin Arlitt, Anirban Mahanti, Characterizing the file hosting ecosystem: A view from the edge, Perform. Eval. 68 (11) (2011) 1085–1102.
[35] Zied Ben Houidi, Giuseppe Scavo, Samir Ghamri-Doudane, Alessandro Finamore, Stefano Traverso, Marco Mellia, Gold mining in a river of internet content traffic, in: International Workshop on Traffic Monitoring and Analysis, Springer, 2014, pp. 91–103.
[36] Max I. Fomitchev, How google analytics and conventional cookie tracking techniques overestimate unique visitors, in: Proceedings of the 19th International Conference on World Wide Web, 2010, pp. 1093–1094.
[37] Lara D. Catledge, James E. Pitkow, Characterizing browsing strategies in the World-Wide Web, Elsevier Comput. Netw. ISDN Syst. 27 (6) (1995) 1065–1073.
[38] Luca Vassio, Hassan Metwalley, Danilo Giordano, The exploitation of web navigation data: Ethical issues and alternative scenarios, in: Fabrizio D'Ascenzo, Massimo Magni, Alessandra Lazazzara, Stefano Za (Eds.), Blurring the Boundaries Through Digital Innovation, Springer International Publishing, Cham, 2016, pp. 119–129.
[39] Jinliang Fan, Jun Xu, Mostafa H. Ammar, Crypto-PAn: Cryptography-based prefix-preserving anonymization, Comput. Netw. 46 (2) (2004).
[40] Paul J. Wright, Ashley K. Randall, Internet pornography exposure and risky sexual behavior among adult males in the united states, Comput. Hum. Behav. 28 (4) (2012) 1410–1416.
[41] Aaron Clauset, Cosma Rohilla Shalizi, M.E.J. Newman, Power-law distributions in empirical data, SIAM Rev. 51 (4) (2009) 661–703.

[42] David Auerbach, Vampire Porn, MindGeek is a cautionary tale of consolidating production and distribution in a single, monopolistic owner, 2014, Slate, https://slate.com/technology/2014/10/mindgeek-porn-monopoly-its-dominance-is-a-cautionary-tale-for-other-industries.html, 2014-10-23.
[43] Lada A. Adamic, Bernardo A. Huberman, Power-law distribution of the world wide web, Science 287 (5461) (2000) 2115.
[44] Zlatka Avramova, Sabine Wittevrongel, Herwig Bruneel, Danny De Vleeschauwer, Analysis and modeling of video popularity evolution in various online video content systems: Power-law versus exponential decay, in: 2009 First International Conference on Evolving Internet, IEEE, 2009, pp. 95–100.
[45] Jacob. Ratkiewicz, Santo Fortunato, Alessandro Flammini, Filippo Menczer, Alessandro Vespignani, Characterizing and modeling the dynamics of online popularity, Phys. Rev. Lett. 105 (15) (2010) 158701.
[46] György Dán, Niklas Carlsson, Power-law revisited: Large scale measurement study of P2P content popularity, in: IPTPS, 2010, p. 12.
[47] GMDT Forecast, Cisco visual networking index: Global mobile data traffic forecast update 2017–2022, Update 2017 (2019) 2022.
[48] Jochen Peter, Patti M. Valkenburg, Adolescents and pornography: A review of 20 years of research, J. Sex Res. 53 (4–5) (2016) 509–531.
[49] Anna Dodson Saikin, Pornography, feminist legal and political debates on, in: The Wiley Blackwell Encyclopedia of Gender and Sexuality Studies, 2016, pp. 1–6.
[50] Marlene Hartmann, The totalizing meritocracy of heterosex: Subjectivity in NoFap, Sexualities (2020) 1363460720932387.

Andrea Morichetta works as a University Assistant at the Distributed Systems Group of the Institute of Information Systems Engineering at the Technical University of Vienna. He received his Doctoral degree in Electrical, Electronics and Communications in January 2020, in Politecnico di Torino. He worked under the supervision of Prof. Marco Mellia, with a grant fully funded by the Big-DAMA project. In 2017 he visited, for a summer internship, Cisco in San Jose, CA. From January 2019 to July 2019, he was a visiting student at AIT, in Vienna, Austria. His research focused on data analysis and machine learning, focusing on unsupervised methodologies applied to networks, with an accent on Security and Traffic Monitoring.

Martino Trevisan received his Ph.D. in 2019 from Politecnico di Torino, Italy. He is currently assistant professor (RTD-A) at the Department of Electronics and Telecommunications in the same university. He has been collaborating in both Industry and European projects and spent six months in Telecom ParisTech, France working on High-Speed Traffic Monitoring during his M.Sc. He visited twice Cisco labs in San Jose in summer 2016 and 2017, as well as AT\&T labs during fall 2018. His research interests are Big Data and Machine Learning applied to Network Measurements and Traffic Monitoring.

Luca Vassio is an Assistant Professor at Politecnico di Torino and member of SmartData@Polito research center for Big Data technologies. He holds a Ph.D. 'cum laude' in Electronics and Communication Engineering and a M.Sc. in Mathematical Engineering. In the last years he was hosted and worked for Bell Labs, MIT, UFMG, EPFL and GE Aviation. He is interested in many fields of data science, from big data analytics problems to the usage of statistical, machine learning and data mining approaches. He is an expert in creating analytical and data-driven models of real phenomena and optimizing performances in different scenarios. His main applications are in the fields of internet measurements, mobility and human behavior.

Julia Krickl holds a Master degree in Global Studies from the University of Vienna, Austria. She currently works as a Junior Researcher for ÖIAT. Her research focus is feminist theory and societal power relations. She received a B.A. in political science from the WWU Muenster, DE, and a B.Sc. from the University of Twente, NL. She worked for the United Nations Office in Vienna and the IAEA, after being a student assistant in Germany for over two years, contributing to the research of the Center for European Gender Studies (ZEUGS). Her research interests lie at the intersection of qualitative and quantitative research, with a focus on gender as power relations and global interdependences.