2016 IEEE 36th International Conference on Distributed Computing Systems

The Internet is For Porn: Measurement and Analysis of Online Adult Traffic

Faraz Ahmed*, M. Zubair Shafiq†, Alex X. Liu*
*Department of Computer Science and Engineering, Michigan State University. {farazah, alexliu}@cse.msu.edu
†Department of Computer Science, The University of Iowa. zubair-shafiq@uiowa.edu

***Abstract*—Adult (or pornographic) websites attract a large number of visitors and account for a substantial fraction of the global Internet traffic. However, little is known about the makeup and characteristics of online adult traffic. In this paper, we present the first large-scale measurement study of online adult traffic using HTTP logs collected from a major commercial content delivery network. Our data set contains approximately 323 terabytes worth of traffic from 80 million users, and includes traffic from several dozen major adult websites and their users in four different continents. We analyze several characteristics of online adult traffic including content and traffic composition, device type composition, temporal dynamics, content popularity, content injection, and user engagement. Our analysis reveals several unique characteristics of online adult traffic. We also analyze implications of our findings on adult content delivery. Our findings suggest several content delivery and cache performance optimizations for adult traffic, e.g., modifications to website design, content delivery, cache placement strategies, and cache storage configurations.**

***Index Terms*—Adult websites; Content delivery; Porn**

I. Introduction

Background. As the saying goes: "The Internet is for porn" [7]. While it is difficult to estimate how much porn content is available on the Internet [27], there are several widely varying estimates available. According to one estimate, there are at least 4 million adult websites on the Internet, which constitute approximately 12% of all websites [3]. It has also been reported that adult websites have more monthly unique visitors than Netflix, Twitter, and Amazon combined [4], [19]. Furthermore, multiple websites in the Alexa's top-500 global list serve adult content [1]. Moreover, a recent measurement study of a tier-1 ISP reported that at least 15% of all mobile video traffic in the Unites States is from adult content providers [9]. Overall, these statistics indicate that online adult content attracts a large number of users and accounts for a substantial fraction of the global Internet traffic. However, despite its significant volume, little is known about the makeup and characteristics of online adult traffic. The understanding of online adult traffic is important for optimizing its content delivery, which involves complex interactions between content delivery networks and ISPs.

Limitation of Prior Art. There is little prior work on dedicated analysis of online adult websites. Only recently, Tyson *et al.* [25], [26] studied behavioral aspects of two popular adult sites: YouPorn and PornHub. These studies reveal several unique characteristics of adult content such as the elasticity of adult content consumption – users consume whatever they find on the front-pages of adult websites. However, the utility of these studies is limited due to three main reasons: (1) they rely on website data obtained by crawling, which is limited in terms of both temporal coverage and granularity; and (2) their analysis cannot distinguish among users because they rely on aggregate view counts of adult videos; and (3) the focus of these studies is on the behavioral and demographic aspects of two specific adult websites, and whether their findings are representative of other adult websites is unclear.

Main Findings. In this paper, we conduct the first large-scale measurement study of online adult traffic using HTTP logs collected from a major commercial content delivery network. The week-long HTTP logs include traffic from several dozen major adult websites and their users in four different continents. Overall, our HTTP logs account for approximately 323 terabytes worth of traffic from 80 million users. Based on these traces, we present some detailed characteristics of five popular adult websites. We choose these websites on the basis of their popularity, e.g., several of these adult websites have been ranked in the global Alexa top-500 list [1]. This selection represents a broad variety of adult websites, ranging from traditional YouTube-style adult video services, adult image sharing services, to adult social networking services.

We provide an in-depth analysis of online adult traffic including its *aggregate*, *content*, and *user* dynamics. Below, we summarize our key findings and their implications on adult content delivery.

1) Aggregate Analysis: Adult traffic primarily comprises of video and image multimedia content. For many popular adult websites, up to 99% traffic volume consists of video and image content. While the majority of users access adult websites from desktop, smartphones and other mobile devices account for a non-trivial fraction of visitors.

2) Content Analysis: Adult content has widely varying sizes: images are generally less than 1 megabyte and videos are on the order of tens of megabytes. Content popularity distributions exhibit the expected skewness. The temporal access patterns of adult websites are unique and different from the typical diurnal access patterns of traditional web content. Our clustering analysis of content popularity reveals groups of objects with diurnal, long-lived, and short-lived temporal access patterns.

1063-6927/16 $31.00 © 2016 IEEE
DOI 10.1109/ICDCS.2016.81

3) User Analysis: User engagement with adult websites is shorter than non-adult websites. However, adult content can be addictive, i.e., some users repeatedly access certain content. For example, at least 10% of video objects have more than 10 requests per unique user.

4) Implications: Our findings have implications on adult website design and content delivery infrastructure management. For example, a vast majority of users do not visit adult websites on smartphones. This finding highlights the need for adult websites to improve their web interfaces and content delivery strategies for mobile devices. As another example, due to their unique diurnal access patterns, it is important to separately account for adult traffic in the traffic forecasting models and network resource allocation. We also find that adult content providers cannot rely on browser cache to store locally popular content because of prevalent use of incognito/private web browsing. Adult content providers can instead optimize content caching performance by customized networked cache configuration. For example, content delivery networks can improve performance and reduce network traffic by pushing copies of popular adult objects to locations closer to their end-users.

Paper Organization. The rest of the paper is organized as follows. Section II reviews prior work on adult traffic analysis. We provide background and details of our data collection methodology in Section III. Section IV discusses measurement and analysis of online adult traffic. We discuss implications of our findings in Section V. Section VI concludes the paper.

II. Related Work

Despite a large number of studies on general web traffic analysis (e.g., [6], [18], [21], [24], [29], [32]), little is known about the makeup and characteristics of online adult traffic. To fill this gap, this paper presents the first in-depth and large-scale characterization of online adult traffic. To provide some context for our study, we review prominent related work below.

Wondracek *et al.* studied the economic and security issues of the online adult industry [28]. In their study, the authors used manual inspection and automated crawling to investigate the characteristics of adult websites. They classified adult websites based on their functionality into the following categories: paysites, link collections, search engines, domain redirector services, keyword-based redirectors, and traffic brokers. They also created two adult web sites from scratch for traffic profiling and vulnerability assessment. They concluded that several prevalent practices in online adult industry are questionable and can be used to conduct malicious activities. This work provides a broad understanding of the economic and security issues of the online adult industry. In contrast, we analyze detailed HTTP access logs of several popular online adult websites to understand their content and user traffic patterns.

More recently, Tyson *et al.* conducted measurement studies of two Web 2.0 adult websites YouPorn and PornHub [25], [26]. In [25], the authors crawled YouPorn to understand user interactions with the website. This measurement study shows that the fraction of user generated content on YouPorn is much smaller as compared to non-adult video websites like YouTube and Vimeo. However, they found that YouPorn has more average views per video than other non-adult video websites. The authors concluded that the video viewing behavior on YouPorn is dependent on two main factors: front page browsing patterns and the number of categories assigned to a particular video. They found that videos with most views are located on the front page of the website and the number of views a video receives is correlated with the number of categories assigned to it.

In [26], Tyson *et al.* crawled PornHub, an adult social networking website, to analyze the demographic makeup of users. The study shows that a majority of profiles on PornHub belong to young males, however female profiles are more popular in terms of receiving more comments and profile views. The authors also found that active profiles have larger social groups and there is positive correlation between content uploaded by a profile and its number of connections. The data analyzed in these studies is collected by periodic crawling of these two adult websites, which is generally limited to meta-data like view counts, ratings, user profile information. In contrast, our work is based on analysis of detailed HTTP access logs of multiple adult websites, which allows us to track individual user content requests at a fine-grained timescale.

Some prior web traffic measurement studies have reported basic statistics of adult traffic in the context of overall traffic makeup. For example, Du *et al.* studied HTTP traffic from Internet kiosks from two developing countries and reported that adult content accounts for less than 1% of total traffic volume [8]. Erman *et al.* studied video traffic in a large 3G cellular network and reported that adult content accounts for approximately 15% of total mobile video traffic [9]. While these studies are useful, they do not specifically analyze adult traffic in terms of its unique characteristics as compared to non-adult traffic. To the best of our knowledge, our work provides the first in-depth analysis of online adult traffic at scale.

III. Data

Most major adult and non-adult content publishers use third-party content delivery networks (CDNs) to efficiently deliver their content to end-users. According to recent estimates, a significant fraction of Internet traffic is served by CDNs [15]. For instance, Akamai alone delivers between 15-30% of all web traffic [2]. A CDN operator typically places content at multiple geographically distributed data centers. A user's request for content (such as a web page, an image, or a video file) is redirected to the closest data center via DNS redirection, anycast, or other CDN-specific methods [11], [20].

For this study, we collected HTTP access logs from multiple data centers of a major commercial CDN for the duration of one week. The HTTP logs include traffic from several dozen major adult websites and their users in four different continents. Overall, our HTTP logs account for more than 323 terabytes worth of traffic from 80 million users. All personally

identifiable information in the HTTP logs (e.g., IP addresses) is anonymized to protect the privacy of end users without affecting the usefulness of our analysis. Each record in our trace includes information about an HTTP request, containing publisher identifier, hashed URL, object file type, object size in bytes, user agent, and the timestamp when the request was received. We use the user agent field to distinguish between different device types, operating systems, and web browsers [10]. Each record in our trace also includes information about the corresponding HTTP response sent by the CDN server, containing the cache status for the requested object. A `HIT` value indicates that the requested object was found in the CDN cache and a `MISS` value indicates that the file does not exist in the CDN cache. We use the HTTP response codes and cache status information to measure caching performance of proprietary CDN caching algorithms.

Through an extensive manual analysis of publisher identifiers, we separated adult content publishers from the rest (dubbed as "non-adult"). We select five popular adult websites in our data set for further in-depth analysis. The names of these websites are anonymized due to business confidentiality agreements. For adult websites, two websites primarily serve YouTube-style adult video content (termed V-1 and V-2), two websites provide image-heavy adult content (termed P-1 and P-2), and one is an adult social networking website (termed S-1).

IV. Measurement & Analysis

In this section, we analyze aggregate statistics of adult websites (e.g., content composition), content dynamics (e.g., popularity, new content injection), and user dynamics (e.g., session length, addiction to adult content).

A. Aggregate Analysis

Content Composition. We first analyze the content composition of adult websites on CDN servers. To this end, we categorize objects based on their file types into video (*e.g.*, FLV, MP4, MPG, AVI, WMV), image (*e.g.*, JPG, PNG, GIF, TIFF, BMP), and other (*e.g.*, text, audio, HTML, CSS, XML, JS). From Figure 1, we note that V-1 primarily stores video objects on the CDN servers, e.g., 98% of all objects are videos. In contrast, V-2 stores a mix of image (84%) and video (15%) objects. V-2 uses a large number of GIF images to show a video summary when users hover the cursor over the video. P-1, P-2, and S-1 mainly store images (99%) on CDN servers.

Traffic Composition. We next analyze the traffic composition of adult websites in terms of the request count and request size during the data collection period. Request count is the total number of requests received from website visitors for all objects. Request size is the total size of objects requested by website visitors for all objects. We again breakdown content across video, image, and other categories respectively. Figure 2(a) shows the traffic composition distributions with respect to their request count. We observe that the majority of traffic on adult websites consists of video and images. Only V-1 has

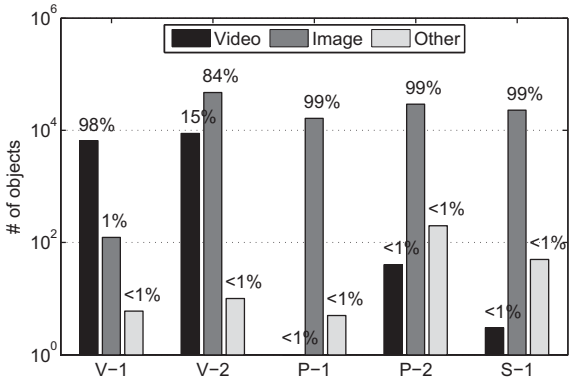
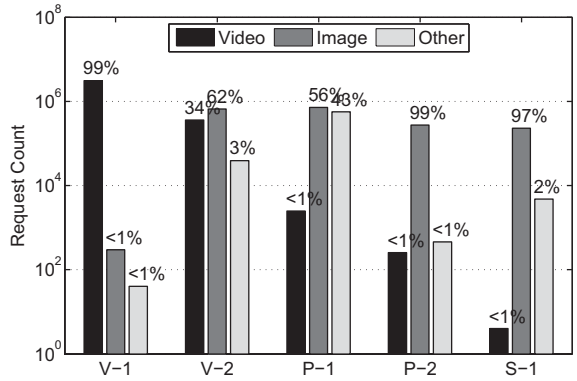
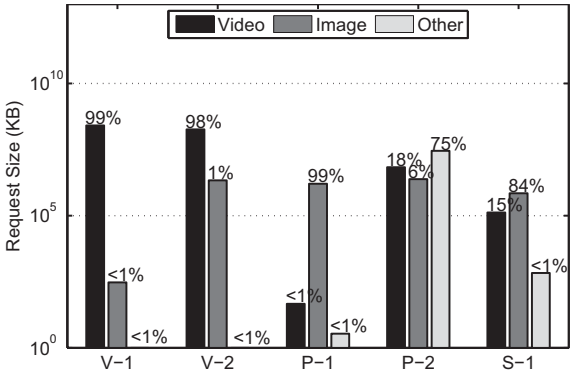


Fig. 1. Content composition of five adult websites. V-1 uses 6.6K objects, V-2 uses 55.6K objects, P-1 uses 16.3K objects, P-2 uses 29.6K objects, and S-1 uses 22.9K objects. We breakdown content into 3 categories: (1) video, (2) image, and (3) other. Other category includes objects that are not classified as video or image.



(a) Request Count



(b) Request Size

Fig. 2. Traffic composition of five adult websites. We note that audio and video multimedia content dominates adult traffic.

significantly more video traffic than other content types – V-1 traffic includes 3.1M requests for video objects. V-2 has smaller percentage of video content traffic as compared to images or other categories. For V-2, 359K requests are for video content whereas 657K requests are for image content. For P-1 and P2, 719K and 175K requests are for images, respectively. For S-1, 231K requests are for images. Figure

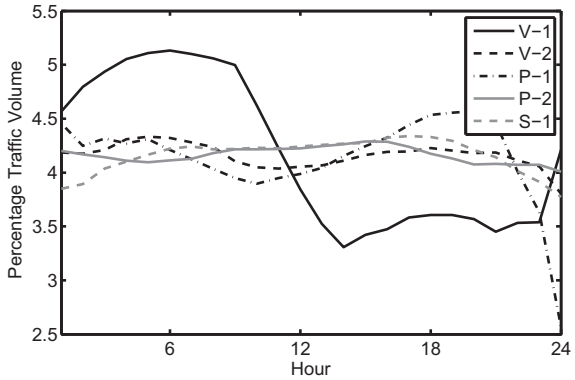


Fig. 3. Hourly traffic volume timeseries of five adult websites.

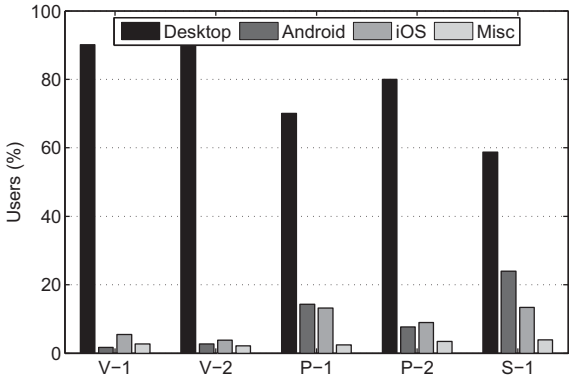


Fig. 4. Device type composition

2(b) shows the traffic composition distributions with respect to their request size. In contrast to Figure 2(a), we note that video content accounts for disproportionately more traffic volume. Since video files are significantly larger than image files, videos tend to dominate the traffic in terms of byte volume. Video traffic in V-1 alone accounts for 258 gigabytes worth of traffic. This traffic mix is composed of mostly multimedia content and it is representative of popular free (ad-driven) and subscription based adult websites [28]. Our findings highlight that the adult content publishers and the respective CDNs need to design and provision their infrastructure to primarily serve multimedia content.

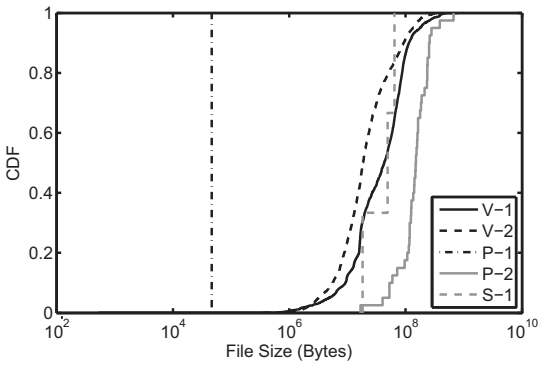
Temporal Access Patterns. We next analyze the temporal access patterns for adult websites. Figure 3 plots the normalized hourly timeseries of traffic volume across the day. We converted the timestamps to local timezones to calculate hourly traffic volumes. Overall, we observe that access patterns of adult websites are *not* typical diurnal patterns. Prior literature (e.g., [21], [24], [29]) reported content access peaks during 7-11 pm and troughs in late night and early morning hours. In contrast, for example, we note that V-1 traffic volume peaks at late-night and early morning hours. This pattern for V-1 is almost opposite to typical diurnal hours reported in prior literature. The temporal access patterns for V-2, P-1, P-2, and S-1 have less pronounced variations than V-1, yet they are different from typical diurnal patterns. The differences between peak access of adult and other content are likely due to the unique nature and viewing preferences for adult content. Thus, it is important for network operators to separately account for adult traffic in the traffic forecasting models and network resource allocation.

Device/OS Usage. The web traffic is known to be gradually shifting from traditional desktop to smartphones and tablets over the last several years [16]. Recall that we extract user agent information from HTTP headers to identify device/OS of a user. We next investigate the device/OS composition of user requests to adult websites. All adult websites discussed in this paper have mobile friendly websites/apps. Figure 4 plots the device distribution of users accessing the adult websites. We observe that the desktop category dominates smartphones (Android and iOS) and miscellaneous (tablets and other mobile devices) categories. For instance, V-2 has more than 95% users accessing content from traditional desktop devices. We observe that image-heavy and social networking websites receive relatively more visitors from smartphone devices as compared to video websites. For instance, more than one-third of users access S-1 from smartphone and miscellaneous device categories. The differences across different adult websites can be partially explained by user preference to view adult content on larger screens. CDNs can customize content delivery (e.g., compression, encoding) depending on different video playback devices. Since a vast majority of users do not visit adult websites on smartphones, our findings highlight the need for adult websites to further improve their web interfaces and content delivery strategies for mobile devices.

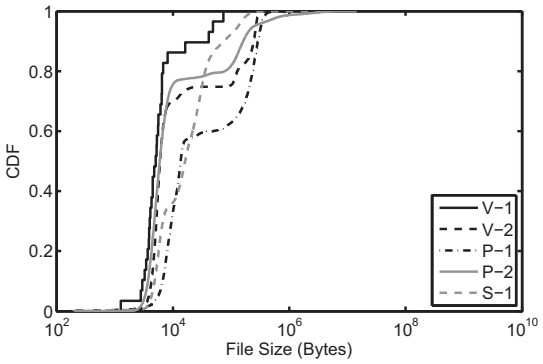
B. Content Dynamics

Content Size. We next investigate content sizes for adult websites. Figure 5 plots the Cumulative Distribution Functions (CDFs) of content sizes. Overall, we observe that content sizes vary in the range of a few kilobytes (KB) to hundreds of megabytes (MB), where majority of requested video objects have sizes greater than 1 MB and image objects are less than 1 MB in size. Figure 5(a) shows the content size distribution of video objects. Video adult websites V-1 and V-2 have a majority of objects larger than 1 MB. P-1 and S-1 have relatively small number of video objects. P-2 has the largest video object sizes. Figure 5(b) shows the content size distribution of image objects. We note that multiple adult websites have bi-modal distributions, indicating thumbnail sized images as well as large images of sizes up to 1 MB. These observations have significant implications for CDN/ISP caching optimization. For instance, ISPs/CDNs can employ separate caching platforms to optimally serve small and large sized objects. The caching platform for small objects can be optimized for high-throughput I/O; whereas, the caching platform for large objects can be optimized for more storage capacity.

Content Popularity. We now investigate object popularity for adult websites. We quantify object popularity in terms

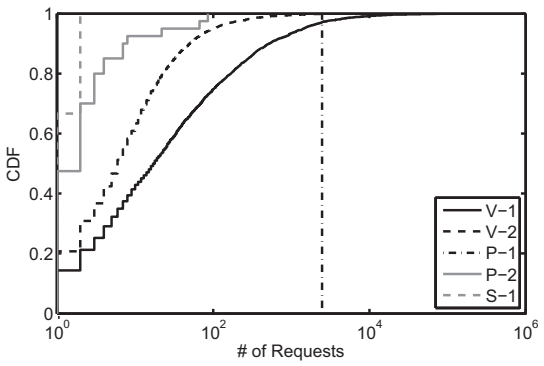


(a) Video

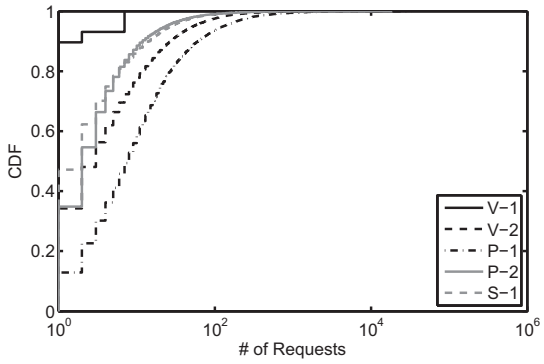


(b) Image

Fig. 5. Content size distributions. We note bi-modal distributions for image objects. Small images are low-resolution thumbnails and large images are high-resolution pictures.



(a) Video



(b) Image

Fig. 6. Content popularity distributions

of request count. Understanding the popularity of objects is important because CDNs typically optimize their caching performance by focusing on popular objects and reduce storage costs by ignoring unpopular and dynamically changing content. Additionally, analyzing the popularity of adult objects can provide insights for identifying similarities and differences between adult objects and non-adult objects. We plot the request count distribution of video and image objects for adult websites in Figure 6. We observe long-tail distributions for all adult websites. This observation indicates that a significant fraction of adult objects are requested infrequently and a small fraction of adult objects are very popular. From a content delivery perspective, this information is useful as CDNs can improve their caching performance by caching heavily requested objects. The long-tailed distribution is similar to those reported for traditional web and video content in prior literature, where a smaller fraction of viral media content is heavily requested. The long-tailed distribution also potentially highlights some social aspects of adult content. In comparison, the large number of requests for typical non-adult objects is mainly because of word-of-mouth content sharing in online social networking websites [22]. We would not typically expect adult content viewers to share adult content on popular social networking websites, though recent work has shown that users use social features in adult websites [26]. We next investigate

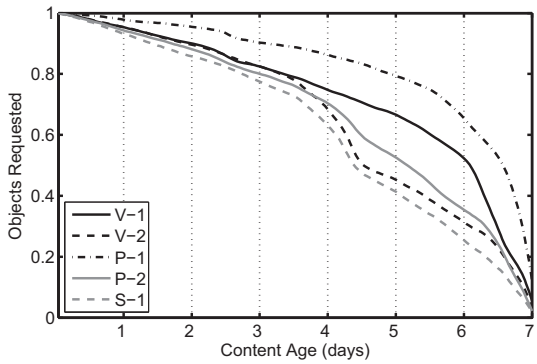
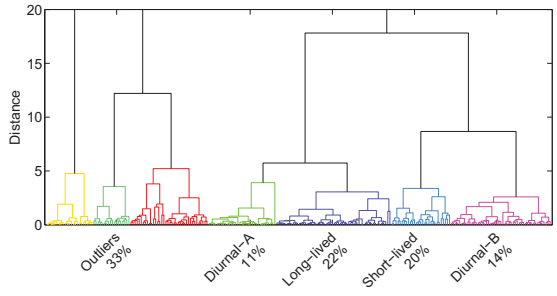


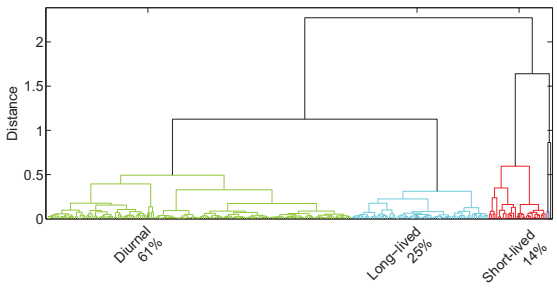
Fig. 7. Fraction of total object requested for adult websites at different ages

unique aspects of adult content popularity by further analyzing individual object request patterns.

Impact of Content Injection and Aging on Popularity. We would expect more requests for objects when they are new, and as the content ages we expect its popularity to decrease accordingly. To understand this phenomenon for adult websites, we plot the fraction of adult objects requested at different ages in Figure 7. The plot shows that a declining fraction of objects are requested as their age increases. In particular, about 20% of objects are not requested after 3 days for most adult websites. Only about 10% of objects are requested throughout the trace duration of one week.



(a) Video V-2



(b) Image P-2

Fig. 8. Content clustering dendrograms of two adult websites.

We further explore the popularity trends of objects by clustering them with respect to their temporal popularity patterns. To identify temporal popularity patterns, we analyze the time series of request count for individual objects. Characterizing the request count time series helps us to assess how fast an adult object reaches its maximum popularity, identifying governing popularity trends of image and video objects, and developing insights for improving the caching performance of CDNs.

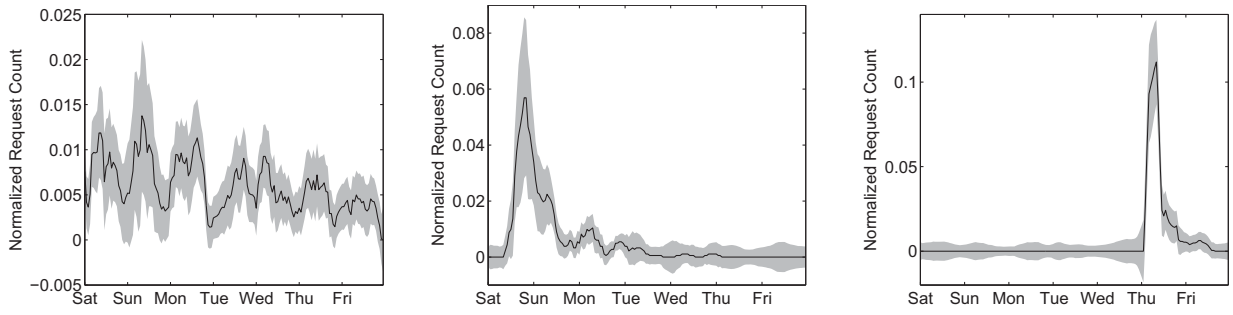
We identify distinct popularity trends by measuring the similarities in shape between the normalized request count time series. We use Dynamic Time Warping (DTW) to compute similarity between two request count time series [17]. DTW uses a dynamic programming approach to obtain a minimum distance alignment between two time series. More specifically, for two time series T_1 and T_2, DTW obtains an optimal alignment between T_1 and T_2 by warping the time dimension of T_1 and T_2. DTW achieves an optimal alignment between two time series by obtaining a non-linear mapping of points on one time series to points of the second time series. This non-linear mapping is also known as the warping path. More specifically, for $T_1 = (a_1, a_2, ..., a_N)$ and $T_2 = (b_1, b_2, ..., b_M)$ of length N and M respectively, a warping path $w = (w_1, w_2, ..., w_L)$ defines an alignment between T_1 and T_2, where $w_1 = (a_1, b_1)$ defines the mapping of element $a_1 \in T_1$ to $b_1 \in T_2$. A cost function is used to compute the optimality of a warping path. Large cost values indicate low similarity in shapes of two time series and small cost values indicate high similarity. The total cost of the a warping path w is defined as:

$$C_w = \sum_{i=1}^{L} c(a_i, b_i)$$

Intuitively, the cost function c is defined as the area between the time warped time series. Using a dynamic programming approach, DTW computes all possible sets of mappings (warping paths) between two time series. The optimal warping path is the path w' that has the minimum total cost among all possible warping paths. The total cost of the optimal warping path is defined as the DTW distance. We use the DTW distance as a metric for quantifying the similarity between two request count time series. We compute pairwise DTW distances for all request count time series and obtain a similarity matrix. We then use the pairwise DTW distance matrix to obtain hierarchical clusters for the request count time series. We use agglomerative hierarchical clustering to obtain dendrogram for each adult website.

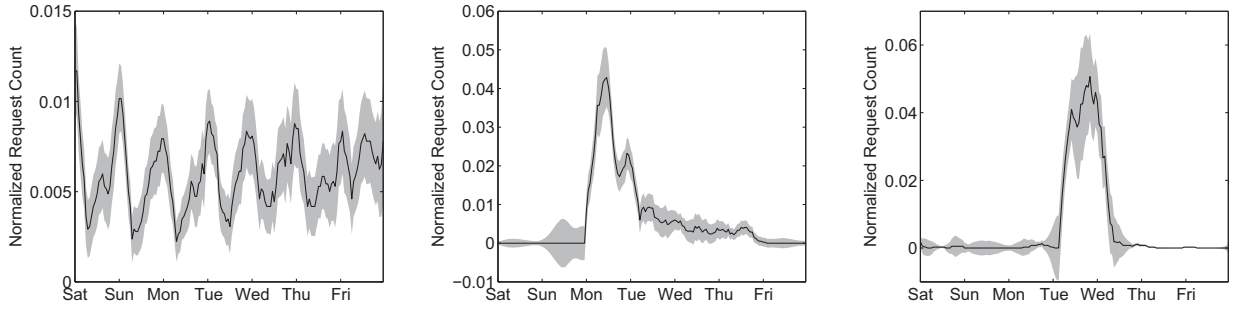
Figures 8(a) and (b) show two example dendrograms of video and image objects for V-2 and P-2, respectively. The x-axis of a dendrogram shows the cluster labels and their memberships and the y-axis represents the DTW similarity metric. We observe four dominant popularity trends in our clustering analysis: `diurnal`, `long-lived`, and `short-lived`. We also observe that some objects in V-2 and P-2 websites have other popularity trends that cannot be neatly categorized as any of the aforementioned categories. We categorize these objects as outliers.

To visualize the unique popularity trends, we next identify a representative sample object from each cluster and plot its normalized request count time series with point-wise standard deviations. To identify the representative sample object for each cluster, we identify its *medoid*, where a medoid is defined as the most centrally located point of a cluster [14]. We calculate point-wise standard deviations by looking at the normalized request count timeseries of all objects in the cluster. We plot the normalized request count time series of mediods of unique clusters in Figure 9. The shaded regions represent the standard deviation of all timeseries from the mean of the cluster and the solid line represent the medoid of the cluster. Figures 9(a) and 10(a) show the medoids for the `diurnal` popularity cluster. Diurnal popularity trends of video objects highlight that certain video objects are requested continuously with regular day/night time variations. Figures 9(b) and 10(b) show the cluster medoids of objects with `long-lived` popularity trend. These objects request count reaches its maximum popularity within the first day of their injection. Their request count decays in a diurnal fashion and completely dies down after a few days. Figures 9(c) and 10(c) show the cluster medoids of objects with `short-lived` popularity trend. The request count of these objects reaches its maximum during the first day of its arrival but decays sharply and completely dies down within a few hours. Our further analysis reveals that video objects with `diurnal` trends are smaller in size as compared video objects with `long-lived`



(a) Diurnal-A (b) Long-lived (c) Short-lived

Fig. 9. Cluster medoids for V-2 adult website



(a) Diurnal (b) Long-lived (c) Short-lived

Fig. 10. Cluster medoids for P-2 adult website

and short-lived trends. The long-lived video objects have the largest size followed by short-lived video objects.

Overall, our analysis reveals that objects served by adult websites have diverse popularity trends. Each website has a different composition of these popularity trends. A large fraction of image and video objects have diurnal request patterns. There are two possible explanations for diurnal patterns. First, prior work [25] suggests that users discover content on adult websites through front page browsing. Objects with diurnal patterns are most likely image and video objects displayed on the front page of a website, and users always access these objects when they visit the website. Second, we note that the potentially addictive nature of adult content could drive users to access specific content repeatedly, similar to non-adult media content.

CDNs can utilize this information to optimize cache control by re-validating diurnal objects less frequently and other objects more frequently, for example, hourly for objects with short-lived access patterns and daily for objects with long-lived access patterns. This can also be achieved by setting longer expire times for objects with diurnal and long-lived access patterns. Furthermore, CDNs can reduce network traffic by pushing copies of objects with diurnal and long-lived request patterns to locations closer to end-users.

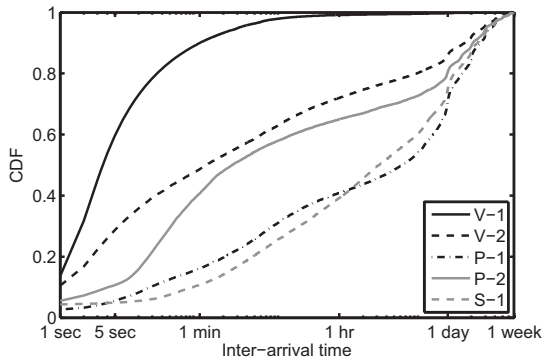


Fig. 11. User request inter-arrival time distributions

C. User Dynamics

User Request Inter-arrival Time. We characterize the user request arrival process in terms of its Inter-arrival time (IAT) distribution. Figure 11 plots the user request IAT distributions for all adult websites. Comparing different adult websites, we observe that video adult websites have shorter request IATs as compared to image-heavy adult websites. For video objects in adult websites, the median request IAT is less than 10 minutes, whereas it is more that 1 hour for image-heavy adult websites. We later use these observations for estimating user session lengths.

User Session Length. A key metric from the perspective of content publishers and CDNs is user engagement [23], which is typically quantified in terms of website bounce time [12].

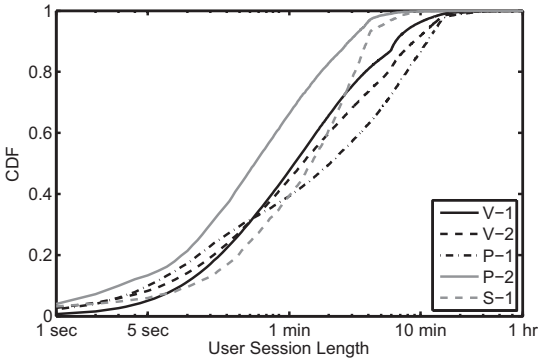
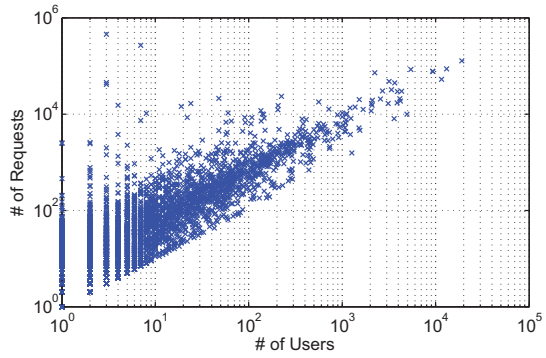
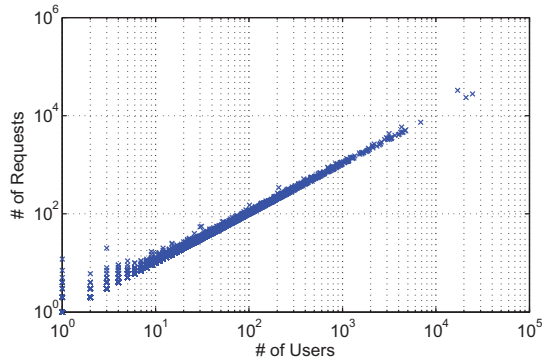


Fig. 12. User session length distributions



(a) V-1

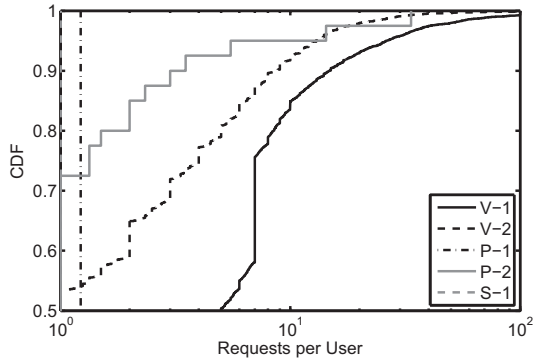


(b) P-1

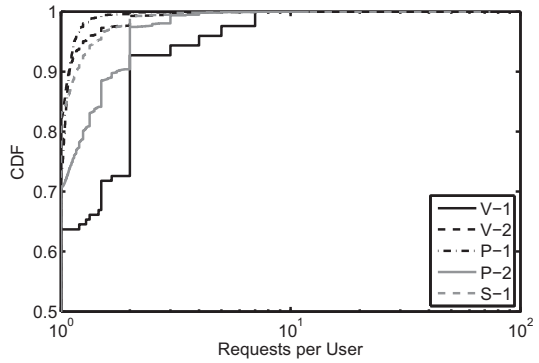
Fig. 13. Repeated access of objects

From the network-side logs, we can estimate user engagement in terms of user session length,[1] where a session consists of consecutive user requests within a timeout interval. We set the timeout value for user sessions at 10 minutes based on our earlier analysis of user request IAT distributions. We plot user session length distributions for adult websites in Figure 12. We observe that the median session lengths for most adult websites are around one minute. Our findings indicate that user engagement for adult content consists of relatively short-lived sessions as compared to non-adult content. We further verified our observations using the engagement statistics reported by

[1]It is noteworthy that the user session length is a strictly lower-bound of traditional bounce time because we cannot tell how long a user continues to watch the downloaded content from network-side logs.



(a) Video

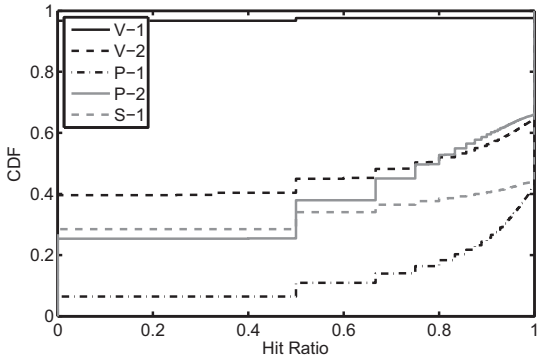


(b) Image

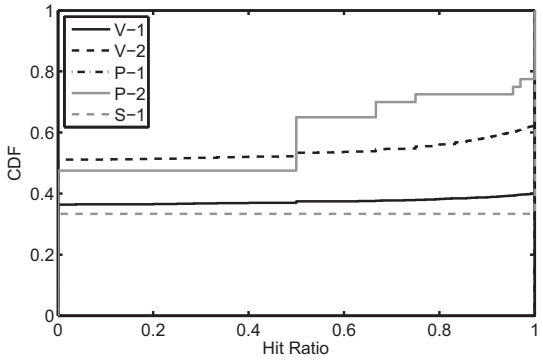
Fig. 14. CDF of repeated content access by users

Alexa. We note that average session lengths for popular non-adult websites tend to be much larger than popular adult websites. For example, the average session length for YouTube [31] is approximately two minutes, whereas the average session length or XVIDEOS [30] is less than one minute.

User Addiction. To further investigate user engagement, we next analyze user addiction. We analyze repeated content accesses by a user to investigate content addiction. For each object, we compute the total number of requests and the total number of unique users who make these requests. Figure 13(a) shows the scatter plot highlighting repeated access of video objects for V-1 adult website. Each data point in the plot represents a distinct video object. We observe that certain video objects are requested by a user multiple times, i.e., data points above the diagonal. In some cases, an object is requested by a large number of times by an individual user. For example, some objects have up to two orders of magnitude more requests than unique users. This observation indicates that some video objects are popular due repeated access by a certain user (i.e., addiction), whereas other video objects are popular due to multiple users accessing the content (i.e., viral). Figure 13(b) shows the scatter plot of repeated access of objects for P-1 adult website. A comparison of repeated access reveal that video objects are more likely to get repeated user requests than image objects. This also highlights that video content is more addictive/engaging as compared to image content.

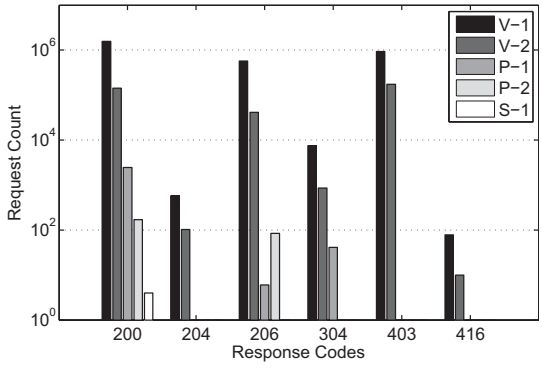


(a) Image

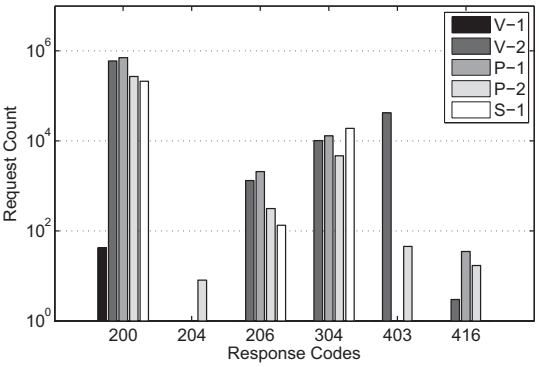


(b) Video

Fig. 15. Hit ratio for image and video objects



(a) Video



(b) Image

Fig. 16. Response codes

To further analyze user addiction to content, we plot the distribution of repeated user access for all objects. Figure 14 plots the CDF of number of requests per user for all adult websites. We note that less than 1% of image objects are requested more than 10 times by a user, whereas at least 10% of video objects have more than 10 requests per unique user. From a CDN's perspective, this information is particularly useful as they can differentiate between objects that are popular only due to requests from multiple users versus those objects that are popular because of repeated accesses by a user. As we discuss later, this information is also useful in optimizing local browser caching and proxy caches deployed by many ISPs. Objects accessed multiple times by a single user or a small number of users should be locally cached closer to end-users.

V. Implications

We discuss potential implications of our measurement and analysis of online adult traffic. We are particularly interested in understanding the impact of different content access patterns on CDN caching. To this end, we analyze the caching performance for adult websites by looking at server-side HTTP response codes and cache hit ratios.

CDN Cache Hit Ratios. We first investigate the cache performance of CDN servers by analyzing server-side cache hit ratios. Figure 15 plots the distributions of cache hit ratios for objects in all adult websites. A hit ratio value of 0 indicates that the requested objects were not present in the CDN cache. A hit ratio value of 1 indicates that all requested objects were served from the CDN cache. Note that the CDN treats video chunks as separate objects for the sake of caching. Comparing image and video objects, we observe that image objects have better overall cache hit ratio than video objects. S-1 has the smallest percentage of objects added to the CDN cache. To understand the cache performance of objects with different popularity, we compute correlations between hit ratio and object popularity. As expected, we find that popular objects tend to have higher hit ratios (more than 0.9 correlation coefficient for all adult websites). Thus, while the distributions in Figure 15 may indicate otherwise, overall CDN cache hit ratios range between 80-90% for different adult websites. It is noteworthy that CDNs often customize cache configuration and performance for individual publishers. Thus, some differences in cache hit ratios may also reflect differences in priorities for different content publishers. Furthermore, customized caching strategies for streaming video content can also be implemented by the CDN.

HTTP Response Codes. We next analyze HTTP response codes for adult websites. Figure 16 shows the number of requests associated with each type of response code. The most common HTTP response codes for both video and image objects include: 200, 206, 304, and 403. We note that a majority of response code are 200. Of particular interest to a CDN operator is the 304 response code, which indicates

that the client's requested object is not modified and the local cache copy is up to date. We note that 304 responses constitute a small fraction of all requests, which indicates the potential for improved localized content caching to improve user performance and reduce traffic load on CDN content replica servers. Despite the cacheability of popular adult objects, 304 response counts are particularly low for adult websites because users are known to browse adult content in incognito/private browsing modes [5]. Web browsers dispose local cache content when users close the incognito/private browser windows. In contrast, note that Facebook reported that more than 65% of their photo requests are served from local browser caches [13]. Unfortunately, while traditional non-adult website can fully utilize browser cache to improve performance and reduce traffic, adult content publishers cannot solely rely on it in designing their content and caching mechanisms.

VI. Conclusion

In this paper, we presented a large scale and in-depth measurement and analysis of online adult traffic. We provide an in-depth analysis of their aggregate, content, and user dynamics. We find that the temporal access patterns of adult website are unique and different from typical diurnal access patterns. While a majority of users access adult websites from desktop, smartphones and other mobile devices account for a non-trivial fraction of visitors. Our clustering analysis of content popularity reveals groups of objects with diurnal, long-lived, and short-lived temporal access patterns. User engagement in adult websites is shorter than non-adult websites; however, adult content can be addictive, i.e., some users repeatedly access certain content. Our findings have implications on adult website design and content delivery infrastructure management. For instance, adult content providers cannot rely on browser cache to store locally popular content because of the prevalent usage of incognito/private web browsing. Content delivery networks can also reduce improve performance and network traffic by pushing copies of popular adult objects to locations closer to the end-users.

Acknowledgements

This material is based in part upon work supported by the National Science Foundation under Grant Number CNS-1464110, the National Natural Science Foundation of China under Grant Numbers 61472184 and 61321491, and the Jiangsu High-level Innovation and Entrepreneurship (Shuangchuang) Program.

References

[1] Alexa Top 500 Global Sites. http://www.alexa.com/topsites.

[2] Facts & Figures - Akamai. http://www.akamai.com/html/about/facts_figures.html.

[3] Internet Filter: Internet Pornography Statistics. http://internet-filter-review.toptenreviews.com/internet-pornography-statistics.html, 2006.

[4] Porn Sites Get More Visitors Each Month Than Netflix, Amazon And Twitter Combined. http://www.huffingtonpost.com/2013/05/03/internet-porn-stats_n_3187682.html, 2013.

[5] G. Aggarwal, E. Bursztein, C. Jackson, and D. Boneh. An analysis of private browsing modes in modern browsers. In *USENIX Security Symposium*, 2010.

[6] X. An and G. Kunzmann. Understanding mobile Internet usage behavior. In *IFIP Networking*, 2014.

[7] AvenueQ. The Internet Is for Porn. https://www.youtube.com/watch?v=LTJvdGcb7Fs.

[8] B. Du and M. D. E. Brewer. Analysis of WWW Traffic in Cambodia and Ghana. In *WWW*, 2006.

[9] J. Erman, A. Gerber, K. Ramakrishnan, S. Sen, and O. Spatscheck. Over The Top Video: The Gorilla in Cellular Networks. In *ACM Internet Measurement Conference (IMC)*, 2011.

[10] R. Fielding, J. Gettys, J. Mogul, H. Frystyk, L. Masinter, P. Leach, and T. Berners-Lee. RFC 2616: Hypertext Transfer Protocol – HTTP/1.1. Technical report, Network Working Group, Internet Engineering Task Force, 1999.

[11] B. Frank, I. Poese, G. Smaragdakis, A. Feldmann, B. M. Maggs, S. Uhlig, V. Aggarwal, and F. Schneider. *Recent Advances in Networking*, chapter Collaboration Opportunities for Content Delivery and Network Infrastructures. ACM SIGCOMM, 2013.

[12] I. Grigorik. Breaking the 1000 ms Time to Glass Mobile Barrier. In *SF HTML5 meetup*, 2013.

[13] Q. Huang, K. Birman, R. van Renesse, W. Lloyd, S. Kumar, and H. C. Li. An Analysis of Facebook Photo Caching. In *SOSP*, 2013.

[14] L. Kaufman and P. J. Rousseeuw. *Finding groups in data: an introduction to cluster analysis*, volume 344. John Wiley & Sons, 2009.

[15] C. Labovitz. First Data on Changing Netflix and CDN Market Share, June 2012. http://www.deepfield.net/2012/06/first-data-on-changing-netflix-and-cdn-market-share/.

[16] A. Lipsman. Major Mobile Milestones in May: Apps Now Drive Half of All Time Spent on Digital. http://www.comscore.com/Insights/Blog/Major-Mobile-Milestones-in-May-Apps-Now-Drive-Half-of-All-Time-Spent-on-Digital, June 2014.

[17] M. Müller. Dynamic Time Warping. *Information retrieval for music and motion*, pages 69–84, 2007.

[18] D. Naboulsi, M. Fiore, S. Ribot, and R. Stanica. Large-scale Mobile Traffic Analysis: a Survey. *IEEE Communications Surveys & Tutorials*, 2015.

[19] O. Ogas and S. Gaddam. *A billion wicked thoughts: What the world's largest experiment reveals about human desire*. Dutton New York, NY, 2011.

[20] M. Pathan and R. Buyya. *Content Delivery Networks, Chapter 2: A Taxonomy of CDNs*. Springer, 2008.

[21] U. Paul, A. P. Subramanian, M. M. Buddhikot, and S. R. Das. Understanding Traffic Dynamics in Cellular Data Networks. In *IEEE Infocom*, 2011.

[22] T. Rodrigues, F. Benevenuto, M. Cha, K. P. Gummadi, and V. Almeida. On Word-of-Mouth Based Discovery of the Web. In *ACM Internet Measurement Conference (IMC)*, 2011.

[23] M. Z. Shafiq, J. Erman, L. Ji, A. X. Liu, J. Pang, and J. Wang. Understanding the Impact of Network Dynamics on Mobile Video User Engagement. In *ACM SIGMETRICS*, 2014.

[24] M. Z. Shafiq, L. Ji, A. X. Liu, and J. Wang. Characterizing and Modeling Internet Traffic Dynamics of Cellular Devices. In *ACM SIGMETRICS*, 2011.

[25] G. Tyson, Y. Elkhatib, N. Sastry, and S. Uhlig. Demystifying Porn 2.0: A look into a major adult video streaming website. In *ACM Internet Measurement Conference (IMC)*, pages 417–426. ACM, 2013.

[26] G. Tyson, Y. Elkhatib, N. Sastry, and S. Uhlig. Are People Really Social on Porn 2.0? In *AAAI Conference on Web and Social Media (ICWSM)*, 2015.

[27] M. Ward. Web porn: Just how much is there? http://www.bbc.com/news/technology-23030090, July 2013.

[28] G. Wondracek, T. Holz, C. Platzer, E. Kirda, and C. Kruegel. Is the Internet for Porn? An Insight Into the Online Adult Industry. In *Workshop on the Economics of Information Security (WEIS)*, 2010.

[29] Q. Xu, J. Erman, A. Gerber, Z. Mao, J. Pang, and S. Venkataraman. Identifying diverse usage behaviors of smartphone apps. In *ACM Internet Measurement Conference (IMC)*, 2011.

[30] xvideos.com Site Overview - Alexa. http://www.alexa.com/siteinfo/xvideos.com.

[31] youtube.com Site Overview - Alexa. http://www.alexa.com/siteinfo/youtube.com.

[32] Y. Zhang and A. Arvidsson. Understanding the characteristics of cellular data traffic. *ACM SIGCOMM Computer Communication Review*, 42(4):461–466, 2012.