# Analyzing Aggregate User Behavior on a Large Multi-platform Content Distribution Service

Raushan Raj[1], Adita Kulkarni[2(✉)], Anand Seetharam[1], Arti Ramesh[1], and Antonio A. de A. Rocha[3]

[1] SUNY Binghamton, Binghamton, USA
{rrausha1,aseethar,artir}@binghamton.edu
[2] SUNY Brockport, Brockport, USA
akulkarni@brockport.edu
[3] Fluminense Federal University, Niteroi, Brazil
arocha@ic.uff.br

**Abstract.** In recent years, Video on Demand (VoD) streaming has increased exponentially as a result of reduced streaming costs and higher bandwidth. For retention of consumers, it is crucial for content providers to understand the behavior of their users and continuously improve performance. In this paper, we analyze the user behavior on *Globo.com*, the largest content distribution service in Brazil. We consider 1.4 billion logs spanning a period of four weeks from October 25, 2020 to November 21, 2020. We analyze the user request patterns and the trends in server's response time. We explore metrics such as protocol, status code, cache hits, user agent, content category popularity and geographical distribution of users. We finally investigate the video popularity distribution and trends in size of content downloaded. We observe that the highest number of requests occur between 8 pm and 11 pm. We observe that 57% of requests are served over HTTPS, while significant portion (43%) are still served over HTTP. Our analysis also reveals that nearly 97% of requests result in a cache hit. Additionally, we observe that the video popularity distribution is skewed and follows a power law with 10% of the videos accounting for 87% of the requests.

## 1 Introduction

Video streaming has become extremely popular in recent years and video traffic is expected to account for more than 70% of the total Internet traffic in the upcoming years [7]. Video on Demand (VoD) streaming services such as Netflix, Amazon, Hulu, YouTube and Globo continue to see a huge increase in consumers globally. These content providers generate large amounts of revenue via user subscriptions and advertisements [16], which necessitate good quality of experience for user retention. With the advancements in video streaming such as live video streaming, Ultra High Definition (UHD) or 4K videos, and Augmented Reality/Virtual Reality (AR/VR) video streaming, the user expectations for uninterrupted and high quality video service continue to increase.


W. Bao et al. (Eds.): ADHOCNETS 2021/TridentCom 2021, LNICST 428, pp. 161–172, 2022.
https://doi.org/10.1007/978-3-030-98005-4_12

To effectively manage the exponentially growing content and consumer population as well as to provide high user quality of service while keeping costs to a minimum, it is critical to investigate user behavior on a large content distribution service. To this end, in this paper, we partner with *Globo.com* [2], the largest content distribution service in Brazil (also ranked 1st in Latin America) to analyze and investigate user behavior on its platform. *Globo* is a Brazilian television network that provides online content via *Globo.com*. According to data released by one of *Globo*'s directors [1], they witnessed an increase of 89% in the number of subscribers to *Globoplay*, one of the component of *Globo.com*, in 2020 as compared to 2019. They now stream around 100 million hours of content every month. The main content categories on *Globo.com* are news, sports, entertainment, technology and food.

We collect and analyze around 1.4 billion user requests made to *Globo.com* between October 25, 2020 and November 21, 2020 at server side. We begin our analysis by studying the user request patterns and the time taken by the server to respond to user requests. We then investigate important network-related metrics such as the protocol used, status code, cache hits, user agent, content category popularity and geographical distribution of users. We conclude our study by examining the video popularity distribution and the trends in the size of the content downloaded by users.

Our main findings are summarized as follows:

- By analyzing the traces, we observe that the highest number of requests to *Globo.com* occur at night between 8 pm and 11 pm and the least number of requests occur between 3 am and 8 am. Though expected, this finding is important as it informs the content provider how to provision for peak load. We also observe that the time needed for the server to serve the requests, is the lowest between 3 am and 8 am. Interestingly, we observe that the time required to serve requests is not significantly impacted by the peak load when compared to the rest of the day. We also observe that the request load is least on Saturday followed by Friday. A possible reason is that people socialize more on the weekend with the result that they spend less time on the Internet and *Globo.com*.
- We investigate the performance impact of different network parameters and interestingly, observe that though majority of the requests are served over HTTPS, a sizable portion of requests are still served over HTTP (43%). For improved security, we believe that more requests will transfer over to HTTPS in future. We also observe that roughly 95% of HTTP(S) requests are satisfied with 200 OK message. We also find that most requests (96.5%) result in cache hit at the server, which indicates that the Globo CDNs are caching content effectively.
- Our analysis also reveals that the most popular web browser and operating systems used by the users to watch videos on *Globo.com* are Chrome and Android, respectively. We also find that majority of the videos watched are related to movies or web series (*Globoplay* component of *Globo.com*). Additionally, we observe that 99.93% requests to *Globo.com* occur from Brazil and the majority of the traffic (around 81%) is generated from the five states Ceará, Bahia, Pernambuco, Paraíba and Maranhão. The request distribution within the country can be attributed to the fact that the CDN we collect the data from is located in the northeast of Brazil.

– We explore the video popularity distribution on *Globo.com* and discover that the distribution is skewed and follows the power law where top 10% of videos account for 87% of the total requests while the remaining 90% videos account for just 13% of the requests. A possible reason for this is people share with others when they find some videos to be good and once they watch a certain kind of videos, the recommendation algorithm recommends similar videos to users making less number of content more popular [20].

Our exploration provides characterization of user activity on *Globo.com*, and facilitates improving the platform's service and designing streaming algorithms to enhance the user quality of service.

## 2 Related Work

In this section, we discuss the existing literature on user behavior analysis on video streaming services.

Research in [3,14,17,18] explores user behavior analysis for live video streaming. Two of them analyze the transmission sessions also considering data from *Globo.com*'s server logs. In [18], authors characterize the behavior of mobile users when watching large popular live events in Brazil. In [17], the same authors extend the previous analysis, using data mining techniques, to extract key factors of popular live streaming sessions to understand what factors may impact the quality of the users' experience using mobile devices. [3] analyze QoE and its impact on user engagement for large-scale live video streaming. Liu *et al.* study personalized 360° live video streaming on two commercial platforms, YouTube and Facebook in [14].

[11–13] focuses on analyzing behavior of mobile users towards VoD consumption. [12] characterizes the geographical patterns of a large-scale commercial mobile VoD system, by measuring uniformity and intensity of geographic interests on videos. Authors in [11] analyze users' behavior, video popularity patterns, impact of the connection type and the type of mobile device used using data from a mobile VoD system. In [13], authors analyze viewing behavior of users with respect to three factors—viewing time, user population, and user locality on PPTV (an Internet video provider in China) logs.

Work in [5,6,8] presents user engagement and performance characterization of video streaming services. Authors in [6] analyze the significance of factors such as service quality metrics, network quality metrics, video content and viewer demography in determining viewer engagement and propose personalized models for predicting individual viewer's engagement. In [5] authors characterize use watching time distributions of 1000 most popular videos on PPLive, a commercial Internet VoD system in China. Ghasemi *et al.* present performance characterization of Yahoo's video streaming service in [8].

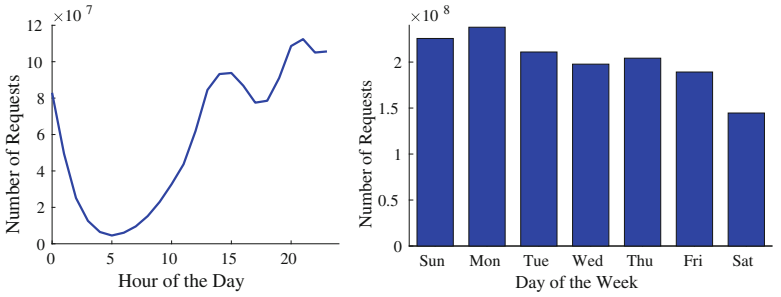
[9,10,15,19] focuses on analyzing the Quality of Experience (QoE) in video streams. [10] studies the relationships between Quality of Service (QoS) and QoE in a session-based Over-The-Top (OTT) video service through a data-driven machine learning approach. Authors explore the use of outlier analysis and clustering as tools for interpreting QoE data of an OTT Video Service in [9]. Authors propose a machine

learning based approach to monitor QoE metrics for encrypted video traffic and present results on YouTube videos in [15]. [19] subjectively assesses QoE during the entire life cycle of video sessions.

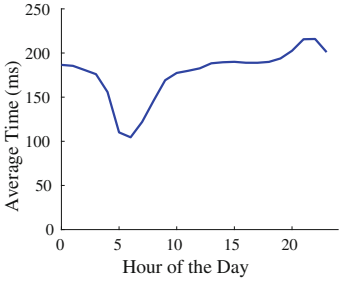
## 3 Data and Problem Statement

### 3.1 Data

In this section, we provide an overview of our VoD streaming dataset obtained from *Globo.com*. *Globo* is the largest television network in Brazil that also offers content over the Internet via its online platform *Globo.com* [2]. For delivering content to its user, the company uses an architecture of multiple Content Delivery Networks (CDNs), located at different cities in Brazil, comprising of multiple servers that cache the most popular content in order to reduce the latency in serving requests to the end users. In this work, we analyze the data collected at one specific CDN, in the northeast of Brazil (state of Ceará). This state (and CDN) is strategically located close to submarine cables connecting Brazil to North America, Europe and Africa. Besides this CDN, *Globo* uses at least others six CDNs to serve data over the Internet.




(a) Requests grouped by hour of the day (b) Requests grouped by day of the week




(c) Response time grouped by hour of the day

**Fig. 1.** Trends in requests sent to *Globo.com* (hourly and weekly basis) and server's response time (hourly basis)

*Globo.com* architecture also uses NGINX web service solution on its HTTP servers. NGINX controls the streaming video service via different HTTP streaming protocols, such as HLS, DASH, MSS, Smoothstreaming, among others. Each NGINX records user session information in log files and sends it to a central repository. We partner with *Globo* guaranteeing access to us to this data repository. Thus, we collect approximately 1.4 billion VoD logs from the service for the analysis presented in this work. Logs contain information of the video requested by a user and server's response to it. As the log is collected at the server, all our analysis is presented from the server's perspective. The logs span four weeks from October 25, 2020 to November 21, 2020. Each log consists of the following fields:

- **Timestamp:** It consists of the date and time when the request was served. The date is logged in YYYY-MM-DD format and the time is logged in HH-MM-SS format.
- **IP address:** It consists of the IP address from where the request was sent.
- **Country Code:** It consists of the postal abbreviation code of the country from where the request occurred.
- **Status Code:** It includes a three-digit numeric code which decides how the user agent (defined in the last point) handles the response.
- **Cache Hit/Miss:** If the requested content is present in the server's cache, it is logged as a hit. If not, it's a miss and the content is retrieved from the backend server.
- **Payload:** It includes the size of the content in bytes that is returned to the user.
- **Response Time:** It consists of time in milliseconds which determines the time elapsed between when the user sent the request and when the request was served. It comprises the time for either finding the video in the server's cache or retrieving it from the backend server and returning it to the user.
- **Video Identification:**This field contains the details of the content requested. It includes the video id, video name and the protocol used.
- **Uniform Resource Locator (URL):** This field contains the web address of the content requested by the user.
- **User-Agent:** It includes the information of the software that renders the web content to the user. This software agent is usually the web browser, media player or a plug in.

### 3.2 Problem Statement

In this paper we adopt a data-driven approach to analyze the underlying patterns in the *Globo.com* dataset to discern key insights related to user trends (e.g., request rate), network performance metrics (e.g., cache hit rate) and video popularity distributions. Our investigation provides a superior understanding of user interactions on *Globo.com*, a unique multi-platform content distribution service and lays the foundation for improving the platform's service and designing streaming algorithms that enhance the user quality of service.

# 4 Results

In this section, we first analyze the users' request patterns and the server's response to them. We then investigate important metrics such as protocol, status code, cache hits, user agent, category popularity and the geographical distribution of users. We then evaluate the video popularity distribution and the trends in the content size returned to the users.

## 4.1 Trends in User Requests

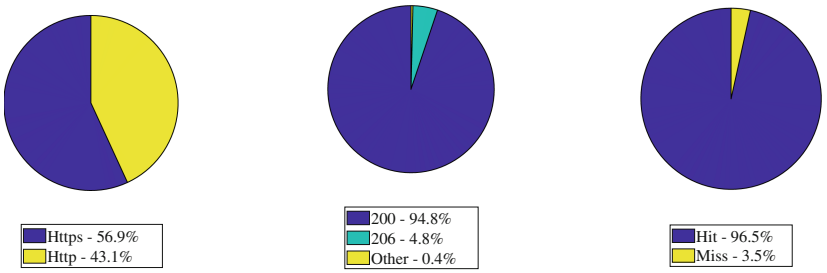
We study the trends in the requests occurring at the server on an hourly and weekly basis. Figure 1a shows the total number of requests occurring at a particular hour of the day during the month. Here, 0 denotes 12 am and 23 denotes 11 pm. We observe that highest number of requests occur at night between 8 pm and 11 pm, after which the number of users accessing the service starts decreasing and the number of requests experience a dip early morning between 3 am and 8 am. This trend is quite natural and corresponds to sleeping patterns of humans. Figure 1b shows the total number of requests occurring at a specific day of the week during the month. We see that highest requests occur on Monday whereas the lowest number of requests occur on Saturday followed by Friday. This is because people prefer spending their weekend nights, in particular Friday night and Saturday night, going out rather than staying at home and watching TV. Analyzing user request patterns is important as it informs the content provider how to provision for peak load.

## 4.2 Trends in Response Time

Response time is the elapsed time between the user requesting a content and the request getting served. It also includes the time taken by the server to search the content in its cache and if not found, fetch it from the back end server. Response time is an important metric to analyze in web surfing and video streaming as low response times, especially while video streaming, could ruin the user experience. Investigating response time is important to improve user engagement and attract more users to use the service. As around 97% requests result in a cache hit at the server (see Sect. 4.3), Fig. 1c shows the average response time of these requests on an hourly basis. We observe that the average time needed to return contents to the users is between 100 ms and 220 ms. We also observe that the average response time is the lowest between 3 am and 8 am. The reason behind this is as we observed in the previous section (Fig. 1a), the server experiences a low number of requests early morning enabling it to serve the requests faster. Interestingly, we observe that the time required to serve requests is not significantly impacted by the peak load when compared to the rest of the day.

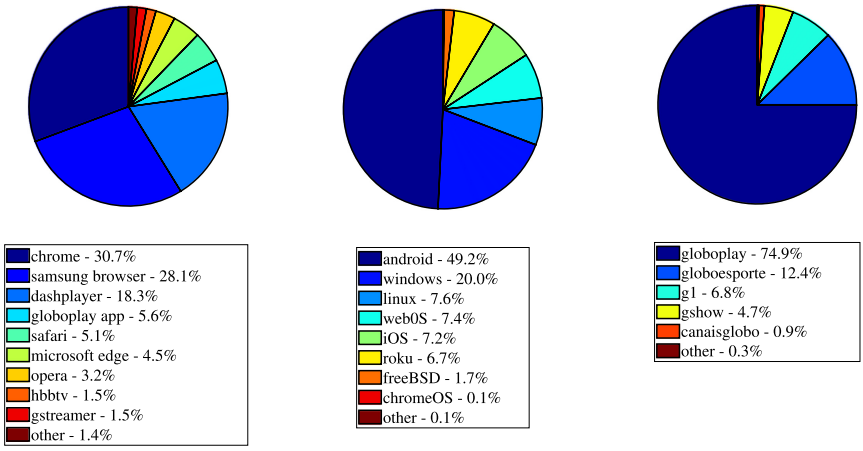
## 4.3 Key Metrics

**Protocol.** VoD services deliver videos to clients over Hypertext Transfer Protocol (HTTP). Hypertext Transfer Protocol Secure (HTTPS) is an extension of HTTP where




(a) Protocol (b) Status Code (c) Cache Hits/Misses

**Fig. 2.** Metrics




(a) Web Browser/Media Player (b) Operating System (c) Content Category

**Fig. 3.** Web browser/media layer, operating system and content category popularity pie charts

the communication protocol is encrypted using Transport Layer Security (TLS). This provides confidentiality, i.e., no one on the network is aware of what the user is watching, and integrity, i.e., no one can alter the video stream. We investigate the protocol used to send requests. Figure 3b shows the percentage of HTTP and HTTPS requests. Interestingly, we observe that around 57% requests are sent over HTTPS whereas a significant percentage of users (43%) still stream videos over HTTP.

**Status Code.** The first line of the HTTP response, called the status line, includes a numeric status code and a textual reason phrase. The way content is retrieved and rendered on web page depends mainly on the status code and thus is an important metric to analyze. Figure 3c shows the percentage of different status codes returned for all the requests. We observe that 94.8% requests are successful and have the status code 200 OK, 4.8% requests are returned only part of the requested resource as they have the

status code 206 Partial Content, while remaining 0.4% requests have 4XX and 5XX class status codes which indicate errors occurring at client and server respectively.

**Cache Hits/Misses.** If the content requested by the user is present in the server cache, it is denoted as a hit. Else, it is a miss and the content needs to be retrieved from the backend server. To reduce the total delay experienced by users getting their requests served, it is important that majority requests incur a cache hit at the server. Figure 3a shows the percentage of requests incurring a hit or a miss. We see that 96.5% requests result in a hit whereas just 3.5% result in a miss, which demonstrates that the *Globo* CDN is able to effectively serve user requests. Further investigation into the 3.5% miss requests can help them understand the reason behind a miss and design approaches to transform such requests into hits.

**User Agent.** A user agent is a software that retrieves and renders web content to the end users. The *user-agent* string in the HTTP request header enables to identify web browsers and media players which act as the user agents for the clients. This information is crucial to the video streaming platforms to provide quality service. From the total logs considered, 31.5% of the requests do not contain information about the user agent. We analyze the remaining 68.5% logs to obtain the most widely used web browser/media player to watch videos and the operating system on which the browser/player runs. Figure 3a shows the most popular web browsers/media players. We observe that about 31% requests occur from Chrome browser, followed by 28% requests from the default browser on Samsung devices and around 18% requests from Dashplayer. Other browsers and media players such as Globoplay App, Safari, Microsoft Edge, Opera, etc. constitute comparatively smaller percentage of requests. Figure 3b shows the percentage of requests from different operating systems (OS). We see that Android is the most popular OS with 49% requests, followed by Windows, Linux, WebOS, iOS, Roku TV OS and FreeBSD, respectively. Other operating systems have minimal requests. This denotes that the majority of the consumers are mobile users.

**Category Popularity.** *Globo.com* offers content in mostly seven different categories—*g1 (journalism), ge (sports), gshow (entertainment), globoplay (TV, web series, and movies), tech (technology), cartola fc (soccer)*, and *receitas (recipes)*. In this subsection, we identify which categories on *Globo.com* are most popular. About 50% of the logs do not contain valid content category information. Thus, we only consider the remaining 50% logs to determine category popularity. Figure 3c shows the pie chart for percentage of requests received for different content categories. We observe that *globoplay (TV, web series, and movies)* is the most popular among users as it incurs the highest percentage of requests, followed by *globoesporte (sports)*, *g1 (news)*, and *gshow (entertainment)*, respectively. Rest of the content receives only around 1.2% of the total requests.

**Geographical Distribution of Users.** Prior research shows that the geographic location of users, especially mobile users, has a significant impact on video popularity [12]. Investigating the geographical distribution enables content providers to better understand regional popularity and plan their service accordingly. We learn from our data that 99.93% requests to *Globo.com* come from Brazil and the remaining 0.07% come from the rest of the world. Figure 4 shows the distribution of user requests in different states of Brazil. As mentioned in Sect. 3, we collect data from the CDN from the state

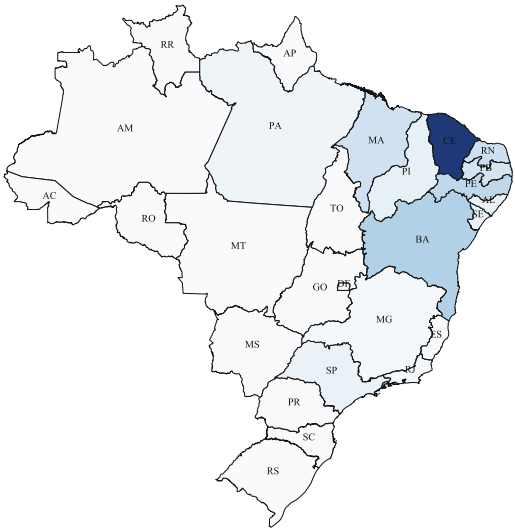
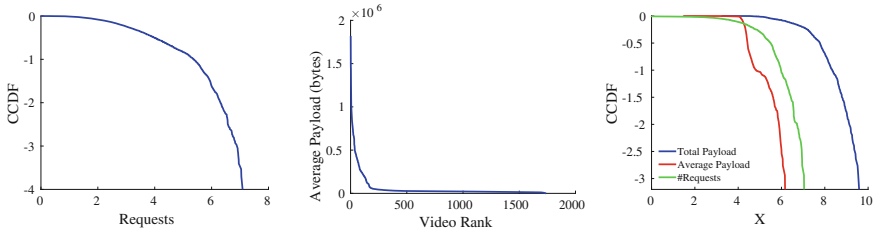



**Fig. 4.** Map of Brazil showing distribution of requests. The five states Ceará, Bahia, Pernambuco, Paraíba and Maranhão contribute majority of the traffic.

of Ceará located in the northeast of Brazil. Therefore, it is understandable that the top five states contributing majority traffic to *Globo.com* in our dataset are Ceará, Bahia, Pernambuco, Paraíba and Maranhão with 41%, 13%, 11%, 7.6% and 7.6% requests, respectively as user requests are generally routed to the geographically closest located CDN. The remaining states contribute significantly less amount of traffic.




(a) Video popularity (log-log scale)

(b) Average content size

(c) CCDF of total number of requests, average payload and total payload (log-log scale)

**Fig. 5.** Video popularity and average content size

### 4.4 Discussion on Video Popularity Distribution and Content Download Size

We investigate the video popularity distribution on *Globo.com*. Figure 5 shows the complementary cumulative distribution function (CCDF) of the video requests. We observe from the data that top 10% videos account for 87% of total requests. The content popularity distribution is skewed and follows the power law. Existing research also indicates that power law is widely prevalent in the real world content popularity distribution [4]. *Globo.com* can leverage these top 10% videos by caching them closer to the end user to reduce latency. Our analysis can also be used to optimize their caching policies and improve the deployment of CDNs. Video popularity also has an important role in video recommendation [20] and therefore, our analysis lays the groundwork to design smarter recommendation algorithms based on the knowledge of popularity changes.

We next analyze the size of the content downloaded from the server. We investigate the total bytes downloaded per video as well as the average bytes downloaded per video considering all user sessions. We obtain a user session as follows—we get the first and the last request received for a video and the user session is the time interval between them. We have two assumptions while considering a user session—i) if two consecutive requests for a video do not occur within a five minute span, we treat them as different sessions. ii) a session cannot be longer than three hours. Both these assumptions are valid because the data that we are investigating is only for VoD and not for live stream videos. So, if a request for the next chunk for the same video does not occur within five minutes, it is highly likely that a user has stopped watching the video. Also, majority of the movies or documentaries are less than three hours. We consider all the sessions for every video requested and get an average of the data (in bytes) returned by the server. Figure 5b shows the average data downloaded for all the videos. We observe that the average size of the data returned for the top 9% videos is greater than 100 MB and up to 1.8 GB. The average size of the content downloaded for all other videos is less than 100 MB.

Figure 5c shows the complementary cumulative distribution function (CCDF) of total number of requests, average bytes and total bytes downloaded per video. We observe that the popularity of the video is correlated to the total bytes downloaded for that video, i.e., videos with higher number of requests have higher number of total bytes downloaded. When we obtain average bytes downloaded considering different user sessions, this does not hold true since the user sessions vary for every video.

## 5 Conclusion

In this paper, we analyzed users' behavior on *Globo.com*, the largest content distribution service in Brazil. We considered user requests made to *Globo.com* over a period of four weeks and investigated the user request patterns and trends in the server's response times. We examined metrics such as protocol, status code, cache hits/misses, user agent, content category popularity and geographical distribution of users. We finally studied the video popularity distribution and size of content downloaded. The findings from this paper can be used by *Globo.com* to make their service more efficient.

## References


1. Deadline.com. https://deadline.com/2021/01/how-brazilian-tv-giant-globo-is-planning-to-compete-with-netflix-amazon-in-the-streaming-war-1234676055
2. Globo.com. https://www.globo.com
3. Ahmed, A., Shafiq, Z., Bedi, H., Khakpour, A.: Suffering from buffering? Detecting QoE impairments in live video streams. In: 2017 IEEE 25th International Conference on Network Protocols (ICNP), pp. 1–10. IEEE (2017)
4. Cha, M., Kwak, H., Rodriguez, P., Ahn, Y.Y., Moon, S.: Analyzing the video popularity characteristics of large-scale user generated content systems. IEEE/ACM Trans. Netw. **17**(5), 1357–1370 (2009)
5. Chen, Y., Liu, Y., Zhang, B., Zhu, W.: On distribution of user movie watching time in a large-scale video streaming system. In: 2014 IEEE International Conference on Communications (ICC), pp. 1825–1830 (2014). https://doi.org/10.1109/ICC.2014.6883588
6. Chen, Y., Chen, Q., Zhang, F., Zhang, Q., Wu, K., Huang, R., Zhou, L.: Understanding viewer engagement of video service in wi-fi network. Comput. Netw. **91**, 101–116 (2015)
7. Forecast, G.: Cisco visual networking index: global mobile data traffic forecast update, 2017–2022. Update **2017**, 2022 (2019)
8. Ghasemi, M., Kanuparthy, P., Mansy, A., Benson, T., Rexford, J.: Performance characterization of a commercial video streaming service. In: Proceedings of the 2016 Internet Measurement Conference, IMC 2016, p. 499–511. Association for Computing Machinery, New York (2016). https://doi.org/10.1145/2987443.2987481
9. Li, W., Spachos, P., Chignell, M., Leon-Garcia, A., Jiang, J., Zucherman, L.: Capturing user behavior in subjective quality assessment of OTT video service. In: 2016 IEEE Global Communications Conference (GLOBECOM), pp. 1–6. IEEE (2016)
10. Li, W., Spachos, P., Chignell, M., Leon-Garcia, A., Zucherman, L., Jiang, J.: A quantitative relationship between application performance metrics and quality of experience for over-the-top video. Comput. Netw. **142**, 194–207 (2018)
11. Li, Z., et al.: Watching videos from everywhere: a study of the PPTV mobile VoD system. In: Proceedings of the 2012 Internet Measurement Conference, pp. 185–198 (2012)
12. Li, Z., Xie, G., Lin, J., Jin, Y., Kaafar, M.A., Salamatian, K.: On the geographic patterns of a large-scale mobile video-on-demand system. In: IEEE INFOCOM 2014 - IEEE Conference on Computer Communications, pp. 397–405 (2014). https://doi.org/10.1109/INFOCOM.2014.6847962
13. Lin, J., Li, Z., Xie, G., Sun, Y., Salamatian, K., Wang, W.: Mobile video popularity distributions and the potential of peer-assisted video delivery. IEEE Commun. Mag. **51**(11), 120–126 (2013). https://doi.org/10.1109/MCOM.2013.6658663
14. Liu, X., Han, B., Qian, F., Varvello, M.: Lime: understanding commercial 360° live video streaming services. In: Proceedings of the 10th ACM Multimedia Systems Conference, MMSys 2019, pp. 154–164. Association for Computing Machinery, New York (2019). https://doi.org/10.1145/3304109.3306220,https://doi.org/10.1145/3304109.3306220
15. Mazhar, M.H., Shafiq, Z.: Real-time video quality of experience monitoring for HTTPS and QUIC. In: IEEE INFOCOM 2018-IEEE Conference on Computer Communications, pp. 1331–1339. IEEE (2018)
16. Research, D.T.: Global OTT TV and Video Forecasts (2020)
17. Correa da Silva, D.V., Velloso, P.B., Rocha, A.A.d.A.: Using data mining techniques to extract key factors in mobile live streaming. In: 2019 IEEE Symposium on Computers and Communications (ISCC), pp. 1–6 (2019). https://doi.org/10.1109/ISCC47284.2019.8969600
18. da Silva, D.V.C., Domingues, G.d.M.B., Velloso, P.B., Antonio, A.D.A.: Analysis of mobile-live-users of a large CDN. In: 2018 IEEE Symposium on Computers and Communications (ISCC), pp. 00946–00951. IEEE (2018)

19. Spachos, P., et al.: Subjective QoE assessment on video service: laboratory controllable approach. In: 2017 IEEE 18th International Symposium on a World of Wireless, Mobile and Multimedia Networks (WoWMoM), pp. 1–9 (2017). https://doi.org/10.1109/WoWMoM.2017.7974323
20. Zhou, R., Khemmarat, S., Gao, L.: The impact of youtube recommendation system on video views. In: Proceedings of the 10th ACM SIGCOMM Conference on Internet measurement, pp. 404–410 (2010)