Measurement of a Large-Scale Short-Video Service Over Mobile and Wireless Networks

Yuming Zhang, Yan Liu, Lingfeng Guo, and Jack Y. B. Lee, *Senior Member, IEEE*

Abstract—Short-video sharing services have seen explosive growth in recent years. Compared to conventional video sharing platforms, these have very different characteristics which are far from well-understood. This work aims at filling the gap by measuring and analyzing detailed *application-level* performance data from a top-10 short video service in China. The application-level data offered detailed and rare insights into many performance metrics of the service, which are otherwise inaccessible to external measurements. The service has a scale of over one billion daily views just for the mobile and wireless segments of the service. Our datasets covered over 22 billion video playbacks, over 100 million video files, served by over 5,000 servers to users across 35 provinces and 13 ISPs in China. We analyzed three aspects of the service: (a) video content characteristics; (b) network analytics; and (c) video streaming analytics. Our results revealed significant differences from conventional video-sharing platforms. These findings will have implications for system designs at all levels. The data also enabled us to conduct an indirect network performance measurement of mobile and wireless network services across China, *as experienced* by the service. These results offer rare insights into mobile and wireless networks' real-world performance in a large country.

Index Terms—Short-video service, video on demand (VoD), data analysis, video content analytics, network analytics, video streaming analytics

1 Introduction

Video sharing is now one of the most popular applications on the Internet. In particular, there has been an explosive growth in short-video sharing (Video on Demand, or VoD) platforms in recent years, with many new media platforms such as Douyin, Tik-Tok, Kuaishou, Kwai, Miaopai, Weishi, and so on. For example, in early 2020, the numbers of daily active users of Douyin and Kuaishou in China were already over 400 and 300 million, respectively [1], [2], [3].

Since most of these short-video services were only recently introduced, their service properties and user behaviors are far from well-understood. Therefore, it is of practical significance to obtain a more comprehensive understanding of these short-video services to enable further advances in their service provisioning.

Previous measurement studies on video-sharing services primarily focused on conventional video-sharing platforms such as YouTube or YouTube-like services [4], [5], [6], [7], [8], [9], [10], [11], [12], [13], [14], [15], [16], [17] where a wide variety of video contents are hosted. By contrast, these new short-video services focus on videos of much shorter durations, often measured in tens of seconds. For example, Gill *et al.* [6] measured YouTube videos to have a median video length of 210 seconds. In contrast, the median video length in Douyin, for example, is a mere 14 seconds [18]. In addition to short video length, many other aspects of short-video service are different, such as user behavior, popularity dynamics, and so on [19]. Therefore, findings from previous studies may no longer apply to this new generation of short-video services. This motivated us to carry out a large-scale measurement study of one of the top-10 short-video services in China (the identity of which shall remain anonymous and henceforth referred to as *the Service*).

We collaborated with a major short video service provider in China, which offered us direct access to *application-level* log data, including streaming and network performance logs that are otherwise inaccessible to external measurements. The Service has a scale of over one billion daily video views just for the mobile and wireless (i.e., Wi-Fi) segments of the provider we studied. Our datasets cover over 22 billion video playback requests, over 100 million distinct video files, served by more than 5000 servers to users from 35 provinces and 13 ISPs in China. We analyzed three key aspects of the Service: (a) video content characteristics; (b) network analytics; and (c) video streaming analytics (c.f. Table 1).

Our results revealed significant differences between short-video and conventional video-sharing platforms [4]–[17]. These findings will have implications for system designs at all levels, ranging from video recommendation algorithms, streaming protocols, to video caching strategies. The Service's scale also enabled us to carry out an indirect network performance measurement of mobile and wireless network services across China, *as experienced* by the Service, incorporated the impacts of all components along the end-to-end data path. The network type (3G, 4G, 5G, and Wi-Fi), geographical location (by province), and ISP used in each video download are

- *Yuming Zhang and Jack Y. B. Lee are with the Department of Information Engineering, Chinese University of Hong Kong, Shatin, NT, Hong Kong SAR. E-mail: zy219@ie.cuhk.edu.hk, jacklee@computer.org.*
- *Yan Liu and Lingfeng Guo are with Cloud ARCH & Platform Department, Tencent, Shenzhen, Guangdong 518000, China. E-mail: rockyanliu@tencent.com, gl016@ie.cuhk.edu.hk.*

Manuscript received 1 May 2021; revised 19 Nov. 2021; accepted 29 Dec. 2021. Date of publication 4 Jan. 2022; date of current version 5 May 2023.
This work was supported in part by General Research Fund under Grant GRF/14206521 from the Hong Kong Research Grant Council.
(Corresponding author: Jack Y. B. Lee.)
Digital Object Identifier no. 10.1109/TMC.2021.3139893

TABLE 1
Comparison of Previous Related Works

		YouTube-like Services	Short-video Services	Others
Video Content Characteristic	Characteristic (Length, Size, Bitrate. . .)	[5-8, 10, 12, 15-17]	[18-19], this work	[13-14, 21-22, 25]
	Consumption (Popularity, Life Span. . .)	[4-7, 9-10, 15-17, 29]	[19-20, 23-24], this work	[13, 21-22]
	Generation (Video Uploaded, Age)	[5-6, 9, 15-17]	[20]	[21-22]
Network Analytics	Flow Characteristic (Size, Duration. . .)	[4-7, 10-12]	This work	[14]
	Throughput	[4, 9-11]	This work	[13-14]
	Round-Trip-Time (RTT)	[7, 10-11]	This work	[28, 30]
Video Streaming Analytics	Play Time, Playback Percentage	[8, 10, 12]	This work	[13-14, 22, 25-26]
	Startup Delay	[7-8, 12]	This work	[13, 25-26]
	Rebuffering	[7-8]	This work	[13, 25-26]
	Adaptive Bitrate	[7-8, 12]		[26]

known (but anonymized), enabling analysis of network performance from multiple angles. These results offer rare insights into mobile and wireless networks' real-world performance as experienced by a large-scale Internet application.

In the following, we summarize the key findings of this study:

- *Conservative provisioning* – The Service's video bitrate is around 930 kbps. Compared to the measured mean network throughput at 30 Mbps, this choice of video bitrate appears to be very conservative. There are several possible reasons, including the lack of bitrate adaptation (Section 4.1 and 7.2), the existence of low-bandwidth users (Section 5.2), the desired to keep rebuffering rate very low (Section 6.2) to maintain user engagement (Section 6.3), and the impact of ISP rate-limiting (Section 7.1).
- *Exceptionally short streaming duration* – The Service's video content has a median video length of merely 22 s (Section 4.1). Moreover, videos were rarely played in full – on average only half of a video was played before it was skipped (Section 6.3). These two factors together means that the streaming session will be exceptionally short. This poses significant challenges to the design of adaptive bitrate (ABR) algorithms as there is very little time for an ABR algorithm to ramp up or converge (Section 7.2). This could be one of the reasons why the Service does not currently adopt ABR in streaming.
- *Rapid popularity evolution* – Video popularity in the Service evolves at a very short time scale (Section 4.3) – minutes versus days in conventional video services such as YouTube. This could impact caching algorithms as those rely on estimation of video popularity to maximize caching efficiency. In addition, we uncovered several factors that could impact the video popularity, including the video upload mode (live versus pre-recorded), time-of-day, and video length. These could potentially be exploited to build new video popularity models for optimizing content management.
- *User behavior* – The Service's users have a very limited attention-span - around 25 s (Section 6.3). For example, only 30.92% of videos were watched in full. The rests were skipped (e.g., screen scrolling [19], [20]) early, with a median play time of just 12 s. We also uncover factors that could impact user viewing behavior. For example, our analysis indicated that the Service's users are very sensitive to playback rebuffering. For example, for the top-10000 videos, a single rebuffering event already reduced the median playback percentage by over 45%. This may very well be one of the reasons for the Service's conservative bitrate choice.
- *Network types* – Our dataset covered all three generations of mobile networks (3G, 4G, and 5G) plus Wi-Fi. An important insight is that Wi-Fi is by far the dominant network type users used to access the Service, accounting for over 80% of all streaming sessions (Section 3.1). This suggests that optimizing Wi-Fi network performance could yield significant gains in service quality. In addition, the results show that Wi-Fi generally outperformed mobile networks, including 5G, in terms of connection time and throughput (Section 5). As a result, the Service performed better over Wi-Fi with substantially lower rebuffering (Section 7.1).
- *Mobile network rate-limiting* – By comparing users' network throughput across different days in a month, we discovered strong evidence of rate-limiting in current mobile networks (Section 7.1). To our knowledge, this is the first empirical evidence showing not just the existence of rate-limiting, but also its impact on throughput and rebuffering performances. This is a new area that warrants further investigation as it could have a significant performance impact on many mobile applications.
- *ABR for short-video streaming* – the Service's adoption of non-adaptive streaming motivated us to explore the potential performance of applying current ABR algorithms to short-video streaming. Our exploratory experiment in Section 7.2 demonstrates that video length has a surprisingly significant impact on the performance of current ABR algorithms. This calls for further investigations to rethink not only the design, but also the metrics to be used in evaluating the performance of short-video streaming services.
- *Empirical models* – The scale of our dataset enables us to model various properties of the Service using known statistical distributions. These models were validated across multiple datasets captured from different periods, so they are representative of the Service, thereby offering an efficient means to support future research, e.g., applied in simulation or mathematical models for

analysis. Details of the model parameters are either included with the figures or documented in the appendices.

The rest of the paper is organized as follows: Section 2 reviews some previous related works; Section 3 introduces our datasets; Section 4, 5, and 6 analyze the video content characteristics, network analytics, and video streaming analytics, respectively; Section 7 discusses two new findings, and Section 8 summarizes the study.

2 Background and Related Work

This section reviews some previous related works on the measurement of the Internet and mobile video services. Table 1 summarizes 27 previous studies on video service measurements. These studies covered one or more of the following aspects: video content [4]–[10], [12]–[25], [29], network analytics [4]–[7], [9]–[14], [28], [30], and video streaming analytics [7]–[8], [10], [12]–[14], [22], [25]–[26].

In terms of the video services being studied, we can divide them into three categories: (i) "YouTube-like Services" covers services such as YouTube and bilibili; (ii) "Short-video Services" covers services such as Douyin and Kuaishou; and (iii) "Others" covers studies analyzing data from a mixture of different video services or a particular vantage point such as a server.

Video content analytics primarily covers video file characteristics, content generation patterns, and content consumption patterns. For example, Cha *et al.* [9] investigated YouTube's video popularity characteristics and measured their popularity evolution over time. Subsequent studies focused on other characteristics, including rating [5], [6], request inter-arrival time [4], file lifespan [15], uploader behavior [27], social networking aspects [5], and advertisements [29]. Besides YouTube, Jia *et al.* [16], [17] investigated bilibili, a service in China similar to YouTube. Their study offered new results on video favorites and viewer-following characteristics, in addition to common video content statistics. Tang *et al.* [22] collected long-term traces of streaming media services to develop MediSyn, a publicly available streaming media workload generator.

More recently, a handful of studies reported measurements of the emerging short-video services. For example, Zhang *et al.* [20] measured video content generation/consumption patterns, users' screen scrolling, and video prefetch strategy for Twitter's Vine service. Zhang *et al.* [23], [24] investigated the video popularity of Kuaishou, and proposed AutoSight, a distributed edge caching system for short-video network. He *et al.* [19] investigated the video file characteristics of Douyin and proposed LiveClip, an adaptive streaming strategy for short-video services.

For network analytics, one important angle is understanding the characteristics (e.g., flow size, flow duration [4-7, 10-12, 14], throughput [4, 9–11, 13–14] and RTT [7, 10–11, 28, 30]) of the network traffic generated by video-sharing services. For example, Zink *et al.* [4] studied YouTube traffic in a campus network and developed a traffic model for simulations. On the other hand, understanding the underlying network performance metrics (e.g., throughput, RTT) could inform the design of streaming applications [28]. For example, Plissonneau *et al.* [10] collected packet-level traces of YouTube and DailyMotion. They analyzed the network RTT, throughput, loss rate, and data wastage – data downloaded but never played. Ghasemi *et al.* [7] studied the bottleneck of Yahoo's video streaming service through analysis of latency (RTT), packet loss, and throughput. Jiang *et al.* [28] studied the network performance of Skype, and proposed VIA to improve Internet telephony call quality, especially under poor network conditions. Zhou *et al.* [30] measured the performance of Taobao-Live, a live video service, and proposed Concerto, a machine-learning-based framework to improve the coordination of application-layer video codec and transport layer protocols. Shafiq *et al.* [14] studied the network dynamics (e.g., signal-to-interference ratio, session inter-arrival time) that could affect video abandonment and developed an empirical model to predict the user's behavior. Despite the many existing works, network characteristics and performance metrics in short-video services have not been reported in the literature thus far.

Finally, for video streaming analytics studies. Previous works studied various streaming performance metrics in conventional video services, e.g., play time and playback percentage [8, 10, 12–14, 22, 25–26], startup delay [7–8, 12–13, 25–26], and rebuffering [7–8, 13, 25–26]. One major focus is on the relation between streaming performance metrics (e.g., startup delay, rebuffering, and bitrate) and user engagement (e.g., video play time, and playback percentage). For example, Chen *et al.* [13] developed an experimental platform with more than 50 self-deployed routers in their university campus and studied the relation between viewer engagement and various system metrics such as rebuffering duration, network throughput, video content, and viewer demography. Dobrian *et al.* [26] investigated the impact of video streaming metrics on video play time using statistical correlation, information gain, and regression methods. Krishnan *et al.* [25] observed the relationship between video streaming metrics and user abandonment/engagement by applying Quasi-Experimental Designs (QED) to uncover causal relationships from observational data. None of the existing studies, however, measured streaming performance and user engagement in short-video services.

3 Datasets

In this section, we describe the way our datasets were collected and summarize their overall properties. Due to space limitations, we focus on the key metrics and findings in the rest of the paper. We refer the readers to Appendix I, which can be found on the Computer Society Digital Library at http://doi.ieeecomputersociety.org/10.1109/TMC.2021.3139893 for additional measurement results and analysis. Moreover, many of the measured metrics could be approximated by mathematical models. We included the model parameters along with the figures for the main results and documented more detailed model parameters in Appendix II, available in the online supplemental material.

3.1 Data Collection

We were provided the *client-side* access logs of the large-scale commercial short-video service. Table 4 summarizes the Service's key properties. The data correspond to one domain name for the Service, which serves short-video

TABLE 2
Features and Notations of the Log

Feature	Notation	Description
Client Timestamp	T_C	The timestamp (ms) when user triggered a (download/playback) event
Client Downloaded Size	B_{DS}	Number of bytes the client downloaded (in a download event)
Transfer Time	D_{TT}	Duration (ms) for network downloading (in a download event)
Video Length	D_{VL}	Length (ms) of video which was played/downloaded. Refer to Section 4.1
Play Time	D_{PB}	Duration (ms) for user spent at this video in a video playback. Refer to Section 6.3
File Name	-	Name of video file which was played (hashed value)
Rebuffering Count	N_{RB}	Number of rebuffering events in a video playback
Rebuffering Duration	D_{RB}	Duration (ms) for rebuffering event in a video playback
Rebuffering Rate	R_{RB}	The ratio of playbacks with at least one rebuffering events. Refer to Section 6.2
Client IP	-	IP address of client (hashed value)
Server IP	-	IP address of server (hashed value)
Client Network Type	-	Network type of client's device, from "3G, 4G, 5G, Wi-Fi"
Video Uploaded Time	-	The time the video was uploaded (accuracy up to hours). Refer to Section 4.2
Startup Delay	D_{SU}	Duration (ms) before video playback. Refer to Section 6.1
DNS Time	D_{DNS}	Duration (ms) for DNS operation.
Connection Time	D_{RTT}	First RTT for http connection. Refer to Section 5.1
Province	-	Client province (hashed value)
ISP	-	Client ISP (hashed value)
Throughput	TP	Throughput for a video playback. Refer to Eq. (4) in Section 5.2
Playback Percentage	R_{PB}	Playback point ratio of a video playback, Refer to Eq. (5) in Section 6.3

TABLE 3
Summary of Key Statistics for the Four Datasets

		Dataset #1	Dataset #2	Dataset #3	Dataset #4
Dataset Information	Logging Date (in 2020)	25^{th}	04^{th}	16^{th}-17^{th}	02^{nd}-10^{th}
	Log File Length	24 h	24 h	48 h	216 h
	Log File Size	584.2 GB	810.3 GB	453.5 GB	949.9 GB
	Log Data Size Per Playback	0.540 KB	0.794 KB	0.503 KB	0.565 KB
Video Content Characteristic	Video Age (Mean/Median)	483/115 h	521/85 h	505/102 h	487/104 h
	Video Length (Mean/Median)	38/20 s	40/22 s	40/23 s	40/22 s
	No. of Playbacks ($x10^9$)	1.08	1.02	0.9	1.68
	No. of Videos ($x10^6$)	62.9	70.5	65.3	108.8
Network Analytics	No. of Segments ($x10^9$)	2.6	4.8	2.1	4.9
	Transfer Time (Mean/Median)	583/358 ms	513/326 ms	478/284 ms	445/273 ms
	Download Size (Mean/Median)	0.92/1MB	0.86/1 MB	0.8/1 MB	0.84/1 MB
	Throughput (Mean/Median)	20.09/17 Mbps	22.89/19 Mbps	27.43/20 Mbps	30.26/21 Mbps
	Connection Time (Mean/Median)	71.54/27 ms	70.63/27 ms	55.68/21 ms	58.96/22 ms
	Persistent Connection Hit Rate	90.02%	94.41%	92.85%	93.17%
Video Streaming Analytics	Play Time (Mean/Median)	25.39/14 s	27.84/13 s	28.78/13 s	27.92/12 s
	Playback Percentage (Mean/Median)	55.43/58%	52.31/47%	52.71/49%	50.93/42%
	Rebuffering Count (Mean)	0.029	0.020	0.019	0.013
	Rebuffering Duration (Mean)	81.9 ms	59.3 ms	51.1 ms	37.0 ms
	Rebuffering Rate	2.08%	1.54%	1.33%	0.96%
	DNS Time (Mean/Median)	47.56/5 ms	30.78/5 ms	64.32/9 ms	47.94/5 ms
	DNS Cache Hit Rate	99.26%	98.79%	99.75%	99.59%
	Startup Delay (Mean/Median)	445/342 ms	446/315 ms	494/294 ms	445/284 ms
	Video Local Replay Cache Hit	0.01%	0.14%	0.01%	0.01%

contents to users using the Service's mobile app on Android and iOS platforms over either Wi-Fi or mobile data networks. The Service has separate servers for its web presence which were not included in this study.

The access log is generated as follows. When a user plays a video, a video playback event is triggered, creating a new log entry. After the user finishes video playback (e.g., returns to the menu or switches to another video), the client app will collect information of the completed playback event and then upload it to the logging server.

The Service does not employ adaptive bitrate streaming such as MPEG-DASH [31], but streams video using a proprietary protocol. It divides each video into segments of around 1 MB in size for delivery to the client over a persistent HTTP connection [33] during playback. The client requests each segment using a separate HTTP request, generating a download event in the log (e.g., download size, transfer time, etc.).

Due to the large volume of traffic and the resultant logs generated, the provider randomly selects and stores 20% of the access logs uploaded. The access logs in this study were all collected in 2020 (Table 3). Table 2 summarizes the main metrics that can be obtained directly or calculated from the logs. Note that the exact month of the log is undisclosed,

TABLE 4
Key Properties of the Service

Property	Details
Streaming infrastructure	Streaming is provisioned using a major CDN provider with over 5000 servers distributed in multiple data centers across China. The exact data center locations were not disclosed.
Video application	A custom mobile application for various smartphone platforms was developed for the Service, including the ability to log and upload streaming performance data to logging servers.
Streaming protocol	Proprietary protocol based on persistent HTTP connection over TCP. Video data are transferred over HTTP in chunks of approximately 1 MB in size.
Video encoding	Average encoding bitrate is around 930 Kbps. Only one video version is available, so streaming is non-adaptive.
Video upload model	Either pre-record or recorded live using camera at three preset video lengths.

but the four datasets are sorted chronologically from #1 to #4, spanning a period of 5 months. To protect users' privacy, all log data were anonymized with each of the fields, namely {Client IP}, {Server IP}, {Filename}, {Province}, and {ISP} replaced by their respective one-way hash values before transferring to us for analysis.

3.2 General Properties

This section presents some general properties of the datasets. We primarily analyzed the data from the video request level as the dataset does not contain information that can identify an individual user. Note that inferring user by (hashed) client IP may not be accurate either as a user's IP may change over time (e.g., when switching between Wi-Fi and mobile network), and ISPs often assign private IPs to their users who then access the Internet via a Network Address Translation (NAT) device [32].

We first measure the distribution of client network type by request in Table 5. Over 80% of requests originated from Wi-Fi users. The rest were split across the three generations of mobile networks, with 4G accounting for most of the requests. This result shows that the Service's user experience is dominated by Wi-Fi, or more specifically, the Service's performance over Wi-Fi (Section 5).

It is worth noting that although 5G deployment is still in its early days, we could already see steady growth over the four datasets, e.g., increasing from <0.01% in dataset #1 to 0.52% in dataset #4 (amounting to 2.7% of all mobile accesses). Simultaneously, 3G requests also decreased progressively as operators phase out 3G services, reaching about the same proportion as 5G after 5 months.

Geographically, Fig. 1 shows the distribution of the requests across all 35 provinces in China. The top three provinces accounted for 30% of all requests. Fig. 2 plots the distribution of the requests across users' ISPs. The top three ISPs accounted for over 93% of all requests. Overall, the above results show that the dataset covered a broad spectrum of requests with respect to network type, ISP, and geographical locations.

TABLE 5
Network Type Distribution

Client Network	Wi-Fi	3G	4G	5G
All datasets	84.87%	0.60%	14.24%	0.29%
#1	81.94%	1.02%	17.05%	<0.01%
#2	90.01%	0.67%	9.32%	<0.01%
#3	84.32%	0.54%	14.83%	0.31%
#4	81.08%	0.58%	17.83%	0.52%

4 Video Content Characteristics

In this section, we analyze the video content characteristics of the Service. These characteristics will be useful in the design and optimization of CDN infrastructures [6] and caching policies [4], [5], [6], [9], or even Peer-to-Peer (P2P) content distribution strategies [4], [5], [9]. Table 3 summarizes the key video content characteristics across all four datasets. Unless stated otherwise, dataset #4 will be used to generate the presented results in the rest of the paper as it is the largest and most recent dataset.

4.1 Video Length and Bitrate

The Service supports two types of video uploads: pre-recorded videos of arbitrary length or live video recorded in real-time in one of three preset video lengths, denoted by L_1, L_2, and L_3 seconds, respectively. In the latter case, the mobile app stops recording after the preset time is up.

The median video lengths in all four datasets are around 22 s, significantly shorter than conventional video sharing services (YouTube at 210 s [5–6, 12]). Fig. 3 plots the video length distribution. We could observe three peaks at L_1, L_2, and L_3, which correspond to the three preset video lengths in the video app's real-time video upload mode. This suggests that users were actively using the feature.

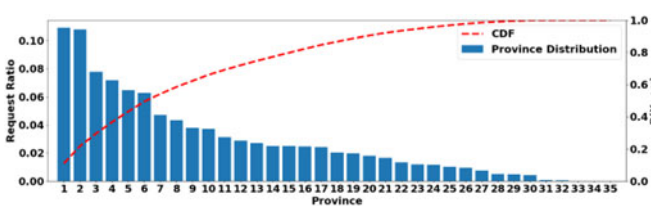

Fig. 1. Geographical distribution (by provinces) of the datasets.

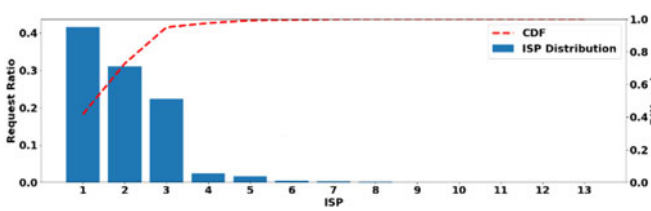

Fig. 2. ISP distribution of the datasets.

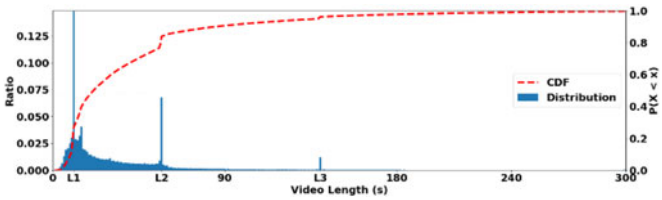

Fig. 3. Video length distribution in dataset #4.

Despite the widespread adoption of adaptive video streaming, most short-video services, including the Service, currently do not employ adaptive bitrate streaming to dynamically adjust the video bitrate within a streaming session [19]. Instead, the Service's videos were encoded with an average bitrate around 930 kbps. We explore the possible reasons for this observation in Section 7.2 by re-examining current ABR algorithms' potential performance when applied to short-video streaming.

4.2 Video Request and Upload Pattern

Next, we study the video request arrival pattern over different times of the day. Fig. 4 plots the per-hour request ratios across two days in dataset #3. The data exhibited clear time-of-day variations in the request pattern, which are consistent across the two days. For example, the peak hour is 21:00-22:00, and the request rate drops rapidly after midnight, reaching the lowest at 02:00-04:00.

The second day had 19.6% more requests than the first day, possibly because the second day was a Saturday. Interestingly, although the first day was a workday, there were still a substantial number of requests *during* office hours. For example, the proportion of requests received between 09:00 to 17:00 was 36.39% on Friday versus 41.35% on Saturday. We conjecture that the very short video length allows users to watch them even during work, e.g., while taking a short break.

One impact of the time-of-day variations in request arrival is on the Service provider's capacity dimensioning. To provide a good user experience, the Service provider must deploy sufficient resources (i.e., servers and network bandwidth) to cope with the peak demand. This means some resources will be idled in the rest of the day.

For example, assuming the Service provisions resources according to the peak demand in the Saturday dataset, then the average resource utilization over the two days is only 31%, and the lowest utilization is only 4.4% at 03:00-04:00. In practice, the utilization will likely be even lower as most, if not all, service providers reserve additional resources as a safety margin to cope with demand spikes.

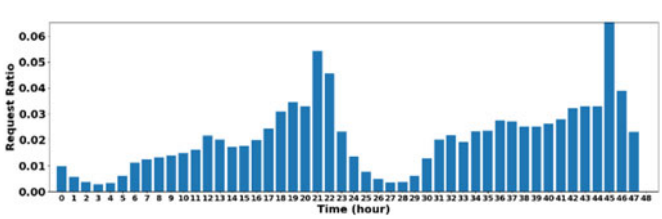

Fig. 4. Time-of-day variation of video playback requests for dataset #3, of which the first day is Friday and the second day is Saturday. Each bar represents the ratio of requests in the time interval [x:00,x+1:00) over the two-day total.

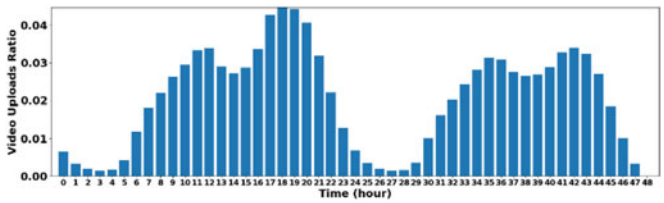

Fig. 5. Time-of-day variations of video uploads for dataset #3.

The measurements in Fig. 4 do offer evidence that the time-of-day variations in video requests could be predictable. By exploiting that, the provider could adapt the resource allocations to match the changing demand to reduce costs and possibly improve performance as well. This is a subject that warrants further research.

Next, we turn our attention to the video upload pattern in Fig. 5 using the same dataset #3. In terms of volume, the ratio of video download to video upload is around 80:1. Again, the upload rate exhibited clear time-of-day variations that are consistent across the two days.

There are, however, some significant differences in the upload pattern. Unlike video requests, which exhibited peak demand during the late evening, the upload pattern exhibited two peaks from 12:00 to 14:00 and 19:00 to 21:00. These two periods coincide with lunch and dinner time which could be one reason behind the higher upload rates.

4.3 Video Popularity

Video popularity is an important metric as it impacts many service designs such as prefetching, caching [34], recommendation, ad-insertion, etc. Previous work [9] measuring YouTube video popularity over a span of several months showed that video popularity generally follows the Pareto rule, i.e., a small subset of videos accounts for a large proportion of views. The popularity versus video rank often can be further modeled by a power-law distribution. However, a recent study by Zhang *et al.* [23], [24] indicated that Zipf's Law might not be applicable to short-video services due to the fast evolution of video popularity.

Zipf's Law – To verify this, we analyzed the popularity of videos in dataset #4. The Pareto rule still applies – 10% of the videos accounted for around 90% of requests. Next, we test if video popularity can still be modeled by power-law distribution, more specifically, the Zipf's Law [35]:

$$F = CR^{-\beta} \tag{1}$$

where F is the access frequency, C is a normalizing constant, R is the video file rank, β is Zipf's coefficient.

Fig. 6 compares the empirical video popularity in dataset #4 to the best-fit Zipf model. We use the coefficient of determination, also known as the R^2 value [36], to measure the

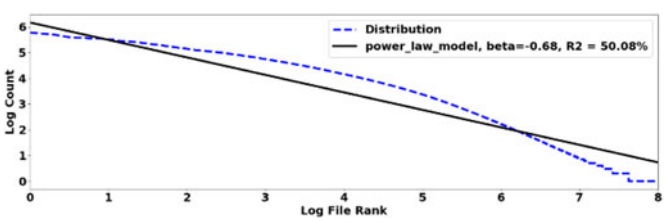

Fig. 6. Measured video popularity versus the Zipf Law.

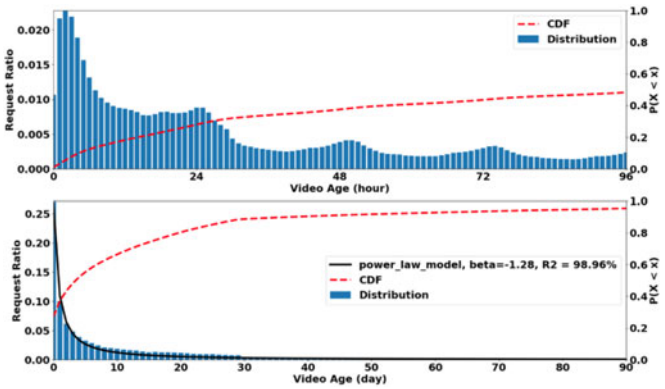

Fig. 7. The ratio of video requests by video age. The top and bottom charts cover videos with ages up to 4 days and 3 months, respectively. The bottom one can be approximated by the power-law model with β = -1.28. The R^2 across 4 datasets are 0.984, 0.962, 0.992, and 0.990.

goodness of model fit where $R^2 = 1$ represents perfect fit, and $R^2 = 0$ is the same as the mean predictor. The R^2 value for the best-fit Zipf model is only 0.5, suggesting the latter is indeed not a good approximation for the video popularity distribution.

Specifically, the high-rank (popular) videos exhibited lower popularity than predicted by the Zipf model. This could be due to the "fetch-at-most-once" property of user-generated content [9], where most videos are watched only once (or a few times in case of very popular videos) by a user.

In contrast, a popular music video on YouTube, for example, will likely be viewed repeatedly by the same user over an extended period of time. Even the most viral short video will be quickly overshadowed by new ones, which is one of the inherent natures of short-video services.

At the other end of the video rank spectrum, we observe truncation of the tail, i.e., their access counts are much lower than predicted by the Zipf model. This observation is consistent with previous works [5], [9], [20], and Cha *et al.* [9] suggested that the truncated tail is due to content filtering. Short video services all have recommendation systems that tend to recommend more popular videos, rendering the unpopular videos hard to be discovered.

Video age – Next, we investigate the evolution of video popularity as *video ages* [22] - defined as the time elapsed since video upload to the time of playback request, to analyze the popularity evolution of a video over its lifetime.

Fig. 7 plots the video requests ratio versus video age in dataset #4 over two timescales (hour and day). In both timescales, video popularity dropped rapidly as a video aged. For example, 30% of all requests were for videos uploaded within the same day, and 90% of all requests were for videos uploaded within 30 days. In comparison, a previous study on Vine by Zhang *et al.* [20] found that 50% of requests were for video within ten days, whereas the same in our dataset was substantially higher at 70%. This rapid popularity decay could impact the performance of caching strategies, e.g., by causing caching pollution [39] in policies based on the most-frequently-used metric. More work is thus warranted to further explore its impact to inform the design of new caching strategies.

The median video age of all requests in dataset #4 is a mere 104 hours. In the by-hour plot, the request ratio exhibited wave-shaped variations over video age with a 24-hour periodicity. This is a result of the time-of-day variations in video requests (Fig. 4) and uploads (Fig. 5).

TABLE 6
Zipf's Model Fit Over Different Time Scales

Duration	Beta (best fit)	R^2
216 hours	0.68	50.08%
24 hours	0.69	64.26%
3 hours	0.68	77.88%
5 minutes	0.60	91.79%

In contrast, the by-day plot exhibited a clear power-law trend that can be approximated by

$$f(x) = cx^{\beta} \tag{2}$$

where x is the video age (in days), f(x) is the ratio, $c = 1$ and β = -1.28 are the power-law coefficients. The power-law model implies that the distribution has a long tail. For example, while most videos have a short video age, a very small, yet not insignificant, number of videos (1.69% to be exact) did exhibit video age longer than 180 days.

The above analysis reveals a subtle yet important parameter in applying Zipf's distribution to model video popularity, namely the *time scale*. The popularity distribution in Fig. 6 was computed from all video requests in dataset #4, which spans nine days. However, as video popularity decays rapidly after just a few hours, it violates Zipf's Law's assumption that the popularity distribution is stationary under the period of study.

To test this hypothesis, we computed in Table 6 the best-fit Zipf model for four different time scales, ranging from 216 hours to 5 minutes. It is clear that the Zipf model's accuracy improves significantly as the time scale shortens. At the 5-min time scale, the best-fit Zipf model has an R^2 value of 91.79%, suggesting a good fit. This shows that the previous observation of non-Zipf behavior in short-video services [23], [24] could be due to the choice of time scale. More importantly, the discovery that video popularity returns to the Zipf model at short time scales opens up a new avenue to exploit it for system optimization such as video replication, caching, and prefetching.

Upload time-of-day – Next, we investigate if the upload time of a video correlates to its eventual popularity. We divided the videos in dataset #4 into 24 groups according to their upload hour-of-day and then plotted their mean request count per video in Fig. 8.

We observe significant variations in video popularity for videos uploaded at different hours of the day. For example, a video uploaded at 17:00-18:00, on average, received 24 playbacks compared to only 11 for a video uploaded at 22:00-23:00. We speculate that this could be caused by two inter-related phenomena. On the one hand, as video ages rapidly with most of its requests generated within the first few hours after upload (c.f. Fig. 7), videos uploaded during 17:00-18:00 could benefit from the subsequent peak request hours (c.f. Fig. 4), resulting in more views. On the other hand, this phenomena could have been exploited by commercial video producers who scheduled their video uploads to reap the most views from the high tide of the request

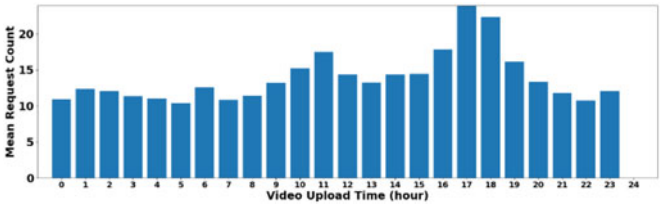

Fig. 8. Impact of upload time-of-day on video popularity.

wave. These uploaders also tend to produce inherently more popular videos, thus skewing the popularity numbers further. A more thorough understanding of these phenomenon will have a far-reaching impact on content and resource management and thus warrant further investigations.

Video length – Finally, we study the impact of video length on video popularity in Fig. 9. The distribution exhibited a concave trend over video length, with video lengths between 60 s to 180 s having higher popularity. The results clearly demonstrate the inherent nature of short-video services – even though longer videos could be uploaded, they are far less likely to be popular because they are not a good fit for the expected use cases (e.g., a quick viewing session at the office).

More surprisingly, the popularities at the three preset video lengths (L_1, L_2, and L_3) for live video upload were significantly lower than their neighbors. We conjecture that there are generally two types of videos. Casual videos captured live by users at the moment of action, using the live upload mode with preset video lengths, and commercial videos that are pre-recorded, edited, and then uploaded using the pre-recorded mode, which has no video length limit. Commercial videos are obviously optimized for popularity, thus attributing to the observed anomalies. This discovery could be low-hanging fruit for optimizing resource management, e.g., by assigning lower priority to cache videos uploaded using the live mode versus the pre-recorded mode.

The above analysis revealed many interesting correlations between various factors and a video's eventual popularity. Such correlations could be exploited to predict the popularity of newly-uploaded video – an important piece of information for video recommendation, content management, resource allocation, etc. Popularity prediction has been studied in the context of conventional streaming services, e.g., the works by Li *et al*. [39] and Goian *et al*. [41]. These previous works may not translate directly to short-video services, however, given the significant differences in the latter's characteristics. We hope the findings in this work will pave the way for further investigations in this area.

5 Network Analytics

In this section, we analyze network analytics such as throughput and connection time derived from the datasets.

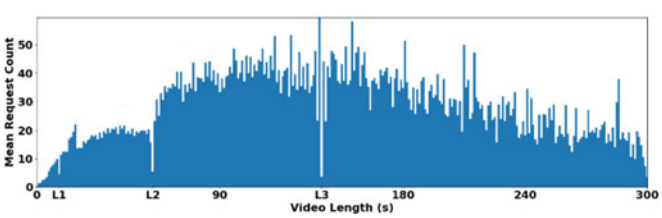

Fig. 9. Impact of video length on video popularity.

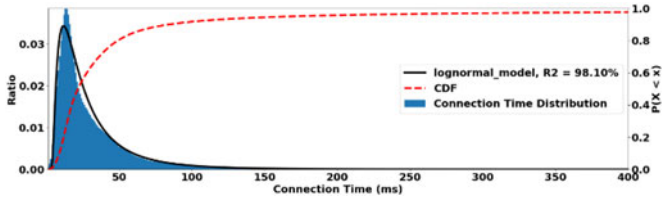

Fig. 10. Connection Time (D_{RTT}) distribution, excluding zero cases and a lognormal approximation with (*s, loc, scale*) = (0.9, 2.95, 19.33). R^2 across 4 datasets are 0.958, 0.947, 0.979, and 0.981.

As opposed to *active* network measurements, which send probing packets and measure their responses [37], the derived data show the network characteristics *as experienced* by the Service, thereby incorporating the impact of all protocol layers as well as the application's inherent behavior.

The datasets covered 35 provinces, 13 different ISPs, and more than 5000 server IPs. Therefore, the results offered a rare opportunity to investigate network characteristics over a country-wide scale. Table 3 summarizes the key network statistics for the four datasets.

5.1 Connection Time

We first consider connection time (D_{RTT}) - defined as the time for the client to establish a TCP connection to the server. With TCP's three-way handshake during connection setup, the connection time reflects the end-to-end round-trip-time (RTT) between the client and the server, plus processing time at both end hosts.

As the Service employs persistent HTTP, only the first HTTP transaction requires connection setup, unless an extended idle period causes the server to timeout the persistent connection. As a result, over 90% of video playbacks have zero connection time due to persistent HTTP. Fig. 10 plots the connection time distribution of dataset #4, excluding the zero cases. We observe that the connection time is relatively short, with a mean/median of 58.96/22 ms. This is because both the Service's clients and servers are located within China, thereby avoiding trans-continent links. For comparison, Langley *et al.* [38] studied the RTT of TCP connections to Google's servers and found that 20% and 10% of RTTs were longer than 150 ms and 300 ms. By contrast, only 5.5% and 2.97% connection times were longer than 150 ms and 300 ms in dataset #4.

The connection time distribution could be approximated by the lognormal model:

$$f(x) = \frac{scale}{\sqrt{2\pi}s(x-loc)} \exp\left(-\frac{\log^2\left(\frac{x-loc}{scale}\right)}{2s^2}\right) \quad (3)$$

where x is the connection time, f(x) is the probability, and s, *loc*, and *scale* are the model parameters. As the connection time is largely comprised of RTT between the client and the server, the model in (3) could be used in simulations for generating random RTT values or used in mathematical models for analysis purposes. In the following, we analyze the correlations between various factors with connection time.

Network types – Intuitively, RTT, and consequently connection time are network-dependent. We segregated the dataset by network types {3G, 4G, 5G, Wi-Fi} and plotted

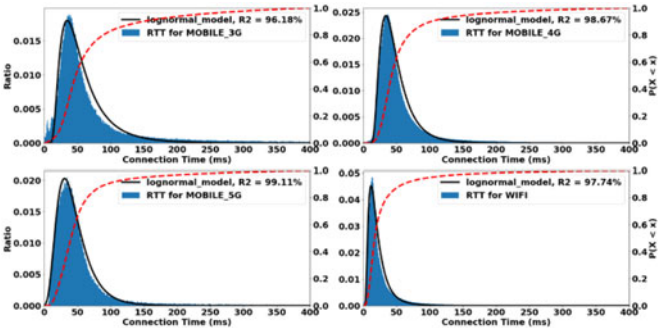

Fig. 11. Comparison of connection time (ms) distribution across 3G, 4G, 5G, and Wi-Fi. R^2 for 4 datasets across different networks are all over 0.96.

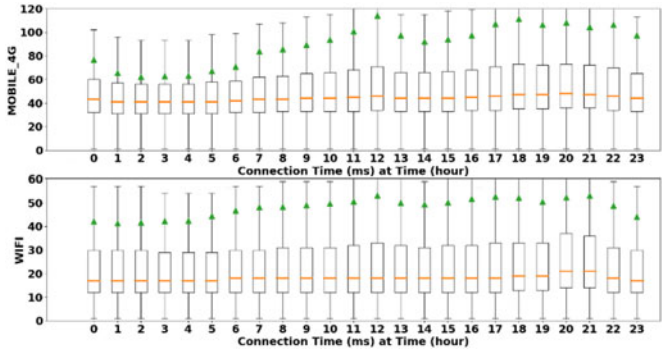

Fig. 13. Impact of time-of-day on connection time. (Boxplot with orange bar and green triangle represent the median and mean).

their respective distribution in Fig. 11. There are two notable observations.

First, connection time generally shortens when going from 3G to 4G and 5G mobile networks, with mean/median values of 151/51 ms, 95/44 ms, and 86/42 ms, respectively. The improved connection time directly translates into a better user experience, especially in short-video services, where short startup delay is far more important than conventional streaming services.

Second, the mean/median connection time for Wi-Fi, at just 49/18 ms, is considerably shorter than even 5G networks. This could be attributed to the early-stage deployment of 5G, where one would expect it to improve over time. Nevertheless, this does show that in addition to reducing mobile data consumption, switching the smartphone from mobile data to Wi-Fi whenever available will also improve application performance.

Time-of-day – We first investigate the relative usage of different network types over different hours of the day in Fig. 12. As expected, most requests were served over 4G and Wi-Fi. What is interesting are the 4G and Wi-Fi curves which are almost mirror images of one another in shape. This shows the impact of Wi-Fi offloading, which is widely adopted by mobile users to reduce their mobile data usage. Nevertheless, despite the variations, the ratio of 4G requests did not drop to near zero even during evenings or midnight, where most users are likely home (and with Wi-Fi access). We hypothesize that some of the users may have subscribed to mobile plans with unlimited data, so there is no need to offload to Wi-Fi. However, our previous connection time analysis (and also throughput analysis in Section 5.2) does show that even if mobile data usage is not a factor, switching over to Wi-Fi will still likely improve application performance.

Next, we analyze the connection time variations over hour of the day in Fig. 13 for 4G and Wi-Fi, respectively. 3G and 5G were omitted due to their very small request ratios. We observe that 4G exhibited more significant time-of-day variations in its mean connection time (green triangle) than Wi-Fi (4G: {62 to 114} ms versus Wi-Fi: {51 to 63} ms). We conjecture that the difference is due to Wi-Fi's generally higher bandwidth which shortens traffic flow duration and can accommodate higher usage bursts. The top two peaks in 4G occurred around lunch (12:00) and dinner time (18:00), possibly due to users being outside the office/home without stable Wi-Fi access, thereby straining the mobile network.

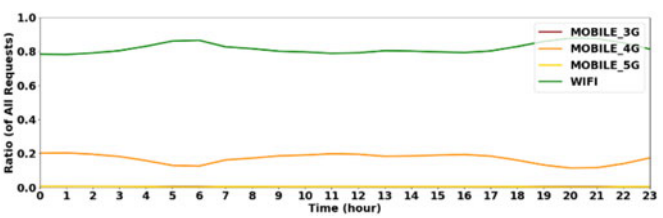

Fig. 12. Network type over time of day of the datasets.

Overall, the above analysis shows that connection time, and likewise RTT, can and do vary substantially across factors such as network type and time-of-day. In particular, despite the fact that both servers and clients are located in China, the mean connection time could still reach over 100 ms in many cases. This is significant for transport layer (i.e., TCP) optimization as long network delay could severely degrade transport protocol performance, especially in high-bandwidth networks [44].

5.2 Throughput

The Service divides a video into 1-MB segments, and the client downloads each segment over a separate HTTP transaction. Each segment download generates a log entry where the download size, denoted by B_{DS}, defined as the number of bytes successfully transferred to the client, is recorded. In addition to segment size, the log also recorded the transfer time D_{TT}, defined as the time to download the video segment.

Knowing the download size and transfer time, one can then compute the throughput in transferring a video segment. Given the Service's scale and reach, this offers a rare opportunity to indirectly measure the end-to-end network throughput performance in a large country. We note that the computed throughput would have incorporated all elements in the end-to-end path, including path bandwidth limit, the dynamics of TCP (i.e., flow and congestion control, loss recovery), the impact of competing flows (sharing a base station or Wi-Fi AP), as well as processing limits of the mobile devices and servers.

In other words, the data measured the end-to-end throughput *as experienced* by the Service and recorded by the client. The available raw bandwidth is likely to be higher, especially over 5G and Wi-Fi, but the recorded throughput reflects what the Service can actually utilize.

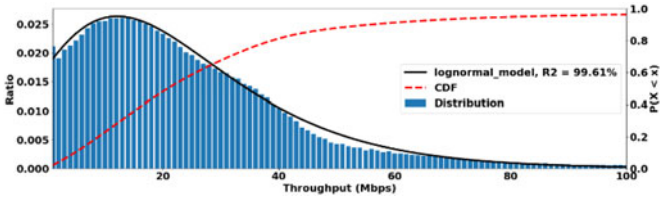

Fig. 14. Throughput distribution and its lognormal approximation with (*s*, *loc*, *scale*) = (0.37, -30.74, 49.16). R^2 across 4 datasets are 0.911, 0.979, 0.981, and 0.996.

Further investigations into the gap between achievable throughput and raw bandwidth will reveal the potential bottlenecks and pave the way to performance optimizations (e.g., transport bottlenecks [44]).

We calculated the mean throughput, denoted by *TP*, for each video session by averaging over all the segment transfers, as follows:

$$TP = \sum_{i \in V} B_{DS_i} \Big/ \sum_{i \in V} D_{TT_i} \tag{4}$$

where V is the set of segments downloaded for the video, B_{DSi} and D_{TTi} are the download size and transfer time of segment i. Note that as the Service employs persistent HTTP, only the first few segment requests would be subject to TCP Slow-Start [44] depending on the bandwidth available. Alternatively, if the user idles for a time exceeding the HTTP keep-alive timeout, then the TCP connection will also need to be re-established with a new Slow-Start phase. Nonetheless, as over 90% of the segment downloads recorded zero connection time, the impact of TCP Slow-Start is relatively small.

We plot in Fig. 14 the overall per-stream throughput distribution, which can be approximated by a lognormal model. In the following, we further analyze the correlations of various factors with throughput.

Network types – We first divide the dataset by network types and plot their respective distributions in Fig. 15. As expected, the throughput generally increases from 3G, 4G, 5G, and finally to Wi-Fi, which has the highest mean throughput. We note that the increase in mean throughput when climbing the network type ladder is somewhat smaller than one may expect.

For example, while 4G generally can offer significantly higher raw bandwidth than 3G, the observed average throughput increased by a mere 21.4% from 11.48 Mbps (3G)

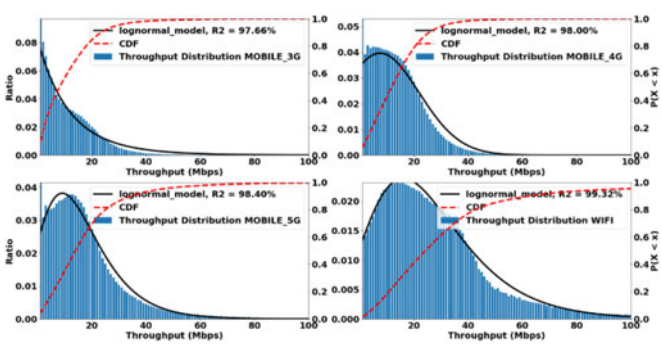

Fig. 15. Comparison of throughput distribution across network types. Note the lognormal models for dataset #4 do not fit other datasets due to mobile network rate limiting (see Section 7.1).

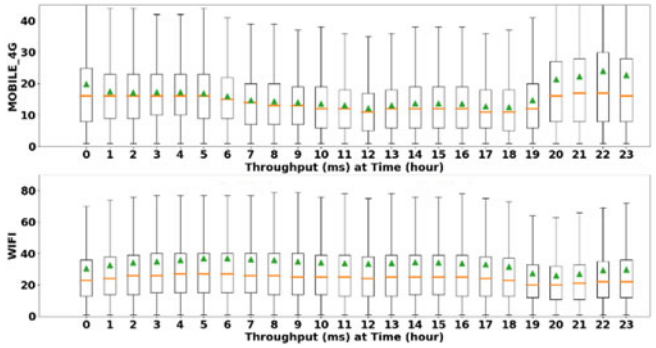

Fig. 16. Impact of time-of-day on throughput.

to 13.94 Mbps (4G). However, 4G did exhibit far fewer low-throughput cases, with only 5.3% having throughput lower than 1 Mbps (slightly higher than the video bitrate of 930 Kbps) compared to 9.7% under 3G.

Similarly, moving from 4G to 5G resulted in a throughput increase of 27.5% from 13.94 Mbps (4G) to 17.78 Mbps (5G) which is lower than what 5G can potentially offer (e.g., over 1 Gbps for 5G [40]). This could be attributed to the early-stage 5G deployment at the time of the study (e.g., only 0.52% of the connections were 5G), so coverage could be limited, resulting in a mean throughput lower than what 5G could have achieved.

Wi-Fi, on the other hand, offers the highest mean throughput at 34.1 Mbps, outperforming the current 5G. The proportion of low-throughput cases (< 1 Mbps) stood at 1.35% for Wi-Fi versus 4.15% for 5G. In addition to Wi-Fi's generally higher bandwidth, its shorter connection time, and hence RTT, could also contribute to higher throughput achieved by TCP.

Time-of-day – Next, we consider the correlations with time-of-day in Fig. 16 for 4G and Wi-Fi connections. Similar to the observations in connection time (Section 5.1), Wi-Fi exhibited more stable throughput throughout the day, ranging from a low of 26 Mbps (20:00-21:00) to a high of 37 Mbps (05:00-06:00). By contrast, 4G ranged from 12 Mbps (18:00-19:00) to 24 Mbps (21:00-22:00).

It is worth noting that the peak hours for 4G centered on lunch and dinner times, where users are more likely to be outside office/home Wi-Fi coverage. The throughput variations of 4G and Wi-Fi are also near mirror-image of one another, demonstrating their complementary nature.

Our further analysis of the four datasets uncovered an often-neglected factor that could also significantly impact network and service performance – mobile network rate-limiting. We present a more detailed discussion of this finding in Section 7.1.

5.3 Throughput-Connection Time Correlation

One key factor impacting throughput performance is TCP itself, as it is used for transporting video data in the Service. Generally, TCP throughput is correlated with link bandwidth, RTT, and loss rate. As connection time is closely related to RTT, we investigate its correlation to mean throughput performance in Fig. 17. It is evident that there are indeed strong correlations between connection time and throughput. Moreover, the relation between connection time and mean/median throughput could be approximated

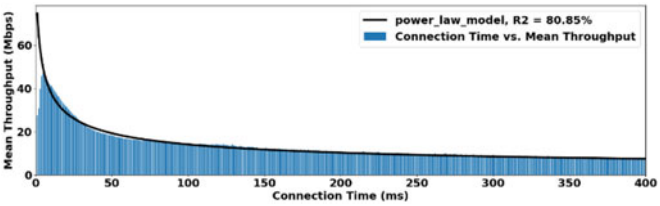

Fig. 17. Dataset #4, Relation between D_{RTT} and mean *TP*. R^2 across 4 datasets are 0.825, 0.757, 0.657, and 0.809.

by the power-law model (see Appendix II, available in the online supplemental material).

We note that for connection time shorter than 5 ms, the power-law model's accuracy decreased significantly. It is not clear why very short connection times exhibited lower than expected throughput. More work is needed to investigate this counter-intuitive phenomenon further.

On the other hand, their correlations suggest that the initial connection time could be exploited to predict the throughput of the connection. This will be very useful to bandwidth-sensitive applications. For example, the predicted throughput could be used to select the video bitrate for the streaming session in non-adaptive streaming platforms or the initial bitrate in adaptive streaming platforms. Again, more work is warranted to explore this finding's potential applications to different Internet applications.

6 Video Streaming Analytics

The application-layer logs in the dataset offered rare details on various performance metrics of the streaming sessions. Instead of inferring or estimating streaming performance from the network analytics [7], [41], the application logs recorded many streaming analytics that are critical to the Service. In the following, we first analyze two key streaming metrics - startup delay and playback rebuffering. Then we analyze the viewing statistics (play time and playback percentage) that reflect user engagement to attempt to quantify the impact of streaming metrics.

6.1 Startup Delay

Startup delay (or startup time) refers to the time from the user clicks a video to the time video playback begins. It is one of the key performance indicators (KPIs) of commercial streaming services as it directly impacts the user experience [26]. It is even more important for short-video services as the video themselves are often only tens of seconds long, and the users typically browse through videos in quick succession looking for interesting ones to watch.

Fig. 18 plots the startup delay distribution and its lognormal approximation. The startup delay is remarkably short,

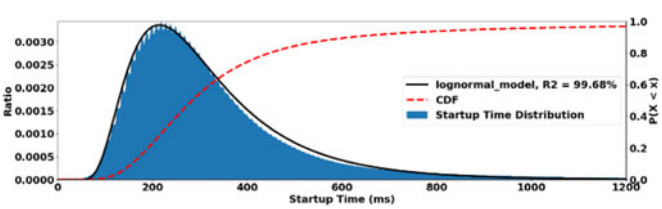

Fig. 18. Startup delay distribution and its lognormal approximation with (*s*, *loc*, *scale*) = (0.56, 33.81, 274.74). R^2 across 4 datasets are 0.908, 0.969, 0.993, and 0.997.

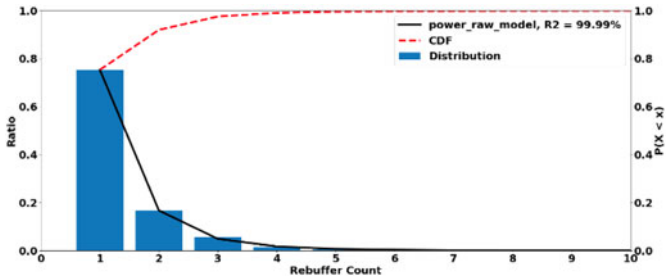

Fig. 19. Rebuffering count distribution for sessions with at least one rebuffer. R^2 across 4 datasets are 0.998, 0.996, 0.999, and 0.999.

with a mean/median of 445/284 ms. In comparison, Finamore *et al*. [12] measured the startup delay in YouTube and found more than 30% longer than 1 s in one of their dataset. By contrast, less than 3% of streams had startup delay longer than 1 s across all four datasets from the Service. The data reflect short-video services' fundamentally different KPIs where startup delay is one of the most important metrics. In retrospect, the Service's choice of moderate video bitrate (∼930 Kbps) and small segment size (∼1 MB) are all deliberate design choices to shorten the startup delay performance that is critical to short-video streaming.

We note that there is a small fraction (0.01%∼0.14%) of zero-startup-delay cases (c.f. last row of *video local replay cache hit rate* in Table 3). This occurs when the same video is played again by the same client using video data that are cached locally. The very low cache hit rate confirmed the fetch-at-most-once property of short video services, as discussed in Section 4.3. Therefore, unlike streaming music services, caching locally in the client is unlikely to reduce server load significantly in short video services.

6.2 Playback Rebuffering

During video playback, rebuffering may occur if the client runs out of video data to sustain continuous playback. The overall rebuffering rate (R_{RB}), defined as the ratio of playbacks with at least one rebuffering event, is quite low in the Service at 0.94%. This is likely due to the conservative video bitrate choice (c.f. Section 4.2). In the following, we focus on the streaming sessions with one or more rebuffering events.

Fig. 19 plots the distribution of rebuffering count, defined as the number of rebuffering events in a video playback. We observe that the majority (76%) of playbacks encountered only one rebuffering event. This is not surprising given the generally short video length. Another way to measure rebuffering is the duration of playback pause - rebuffering duration, plotted in Fig. 20. The results show that when rebuffering occurs, the rebuffering time can be relatively long, e.g., 23% have rebuffering duration $\geq$ 5 s. Given the small video segment size of just 1 MB, this suggests that rebuffering is likely caused by exceptionally poor network conditions, e.g., temporary loss of mobile/Wi-Fi connection, which then takes considerable time to recover.

6.3 Viewing Statistics and User Engagement

Viewing statistics measure metrics of the playback itself, e.g., video play time and playback percentage. These are vital statistics to the Service provider as they reflect the users' experience (also known as user engagement).

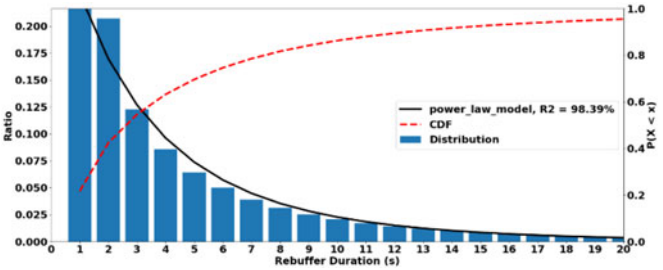

Fig. 20. Rebuffering duration distribution for sessions with at least one rebuffer. R^2 across datasets are 0.962, 0.937, 0.975, and 0.984.

These data are rarely published by service providers as they are proprietary commercial information. Moreover, only the application itself can directly measure these statistics so it is difficult, if not impossible, to measure them externally.

Video Play Time / Playback Percentage – The first metric is video *play time* – measured from the time video playback begins to the time the user terminates playback. Due to their short duration, it is common for short video services to repeatedly play the video in a loop so the play time could exceed the video length. Moreover, it also includes time spent when playback is paused (either explicitly or implicitly due to switching to other applications or answering an incoming call) as well as rebuffering time, if any.

Fig. 21 plots the video play time distribution. As expected, the play time is short, with a mean/median of 27.92/12 s. The video play time distributions of the four datasets can all be approximated by the power-law model, suggesting that the previous assumption of Gaussian distribution (e.g., [19]) for the play time may not apply in actual short video services. To our knowledge, this is the first known result in play time distribution for short video services.

We observe that a substantial proportion of playbacks (31%) have play time shorter than 3 s. These could be attributed to the user's browsing behavior [42], [43], i.e., user finding the initial few seconds of the video uninteresting will terminate it to switch to the next one. This could result in significant data wastage as the rest of the video data downloaded will be discarded [19], [44].

Comparing to the mean video length (e.g., 40 s in dataset #4), the mean play time is even shorter (e.g., 27.92 s). This implies a substantial proportion of the videos was not watched completely. This can be quantified by *playback percentage*, denoted by R_{PB} – defined as the ratio of play time D_{PB} to video length D_{VL}:

$$R_{PB} = \frac{D_{PB}}{D_{VL}} \tag{5}$$

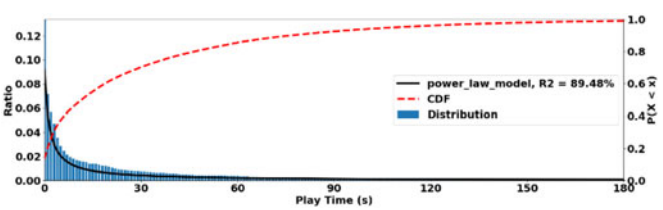

Fig. 21. Play Time distribution, power law model with $\beta = -1.0$. R^2 across datasets are 0.911, 0.901, 0.909, and 0.895.

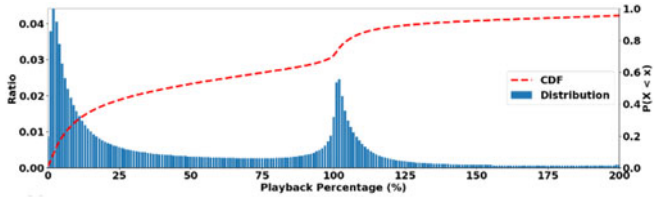

Fig. 22. Playback percentage distribution.

Fig. 22 plots the distribution for the playback percentage. There are two notable observations. First, the playback percentage can exceed 100% due to looped playback, rebuffering, and user interruption. In fact, 30.92% of the playbacks were longer than the video length. Second, there is a significant peak around 100%. The increase when approaching 100% could be attributed to the user recognizing the video is near the end and switch to another video early. In contrast, the peak after 100% is likely due to the looped playback feature that repeats playback of the video after it has reached the end. Noticing the repeat, the user then terminates and switches to another video, resulting in a playback percentage slightly larger than 100%.

The first peak in Fig. 22 suggests that user tends to decide if a video will be interesting early on during the playback session and will terminate and switch quickly if not interested. However, apparently that early decision may not always be correct as well. As a thought experiment, let say a user will spend the first 10% of the video length deciding if a video is interesting. If the video turns out to be interesting to the user, then the user will play at least 90% of its duration. Otherwise the user will terminate and switch early. Using the data in Fig. 22, this translates into a decision correctness rate of 36%. Further analysis of such data could lead to interesting directions for content recommendation, video data prefetching, caching, and so on.

In the following, we further analyze the correlations between playback percentage and four other metrics.

Playback Rebuffering – In addition to startup delay, another important KPI in the Service is playback rebuffering. To exclude the influents of the content itself, we compare the playback percentage of streaming sessions of the same video with different numbers of rebuffering events by dividing them into three groups: (i) no rebuffering; (ii) one rebuffering; and (iii) more than one rebuffering. Table 7 summarizes the playback percentage for the top $N = 100\ldots,100000$ videos by popularity.

Compared to the no-rebuffering group, the playback percentage dropped markedly by over 45% even when there was just a single rebuffering event in the streaming session. A more detailed analysis shows that the drops in playback percentage are generally higher for larger N. For example, playback percentages with just one rebuffering dropped from 57.43% ($N = 100$) and 56.44% ($N = 10000$) to 33.49% ($N = 100$) and 30.67% ($N = 100000$), respectively. A possible explanation for this observation is that users are more likely to endure rebuffering if the content is more interesting (i.e., smaller N).

To our knowledge, this is the first known results quantifying the impact of playback rebuffering on user engagement in short video services. For conventional video services, Dobrian *et al*. [26] did a measurement study and

TABLE 7
Comparison of Playback Percentage for Top-N Videos, With and Without Rebuffering

Top N Videos	$N_{RB} = 0$	$N_{RB} = 1$	$N_{RB} > 1$
100	57.43%	33.49%	35.52%
500	53.88%	30.89%	31.61%
1000	54.41%	31.27%	30.75%
5000	55.61%	30.43%	30.98%
10000	56.44%	30.67%	30.70%

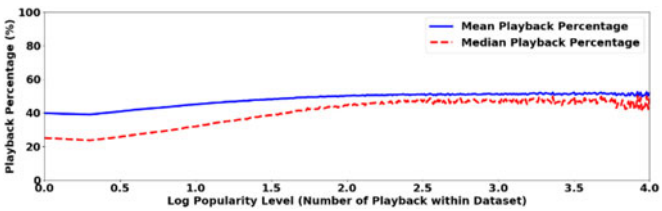

Fig. 23. Playback percentage versus video popularity level.

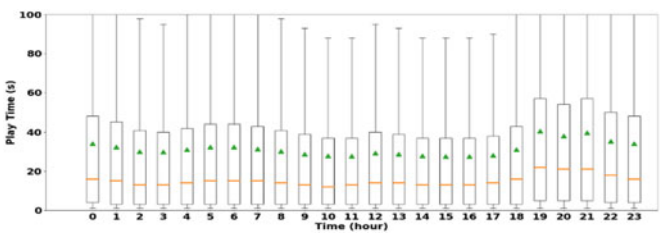

Fig. 24. Play time across time-of-day.

found that a 1% increase in the rebuffering duration could reduce the playback percentage by 1.6% to 4.8% for a 60-min video. As the metric is different, we cannot draw direct quantitative comparison.

Nonetheless, the results from the Service do show that users are very sensitive to rebuffering. Therefore, more works are needed to incorporate the new quantitative results into the design of new QoE metrics that are better suited for short video services.

Video popularity – Next, we investigate the correlation between video popularity and playback percentage. Intuitively, one would expect the two to be positively correlated. The actual results, plotted in Fig. 23, are more complicated. Specifically, playback percentage did increase with video popularity, but only up to around $10^{2.5}$ views where it plateaued. This shows that even for popular videos, a substantial portion of users may still watch only a part of the video before moving on to the next one.

Time-of-day – Another interesting angle is the impact of time-of-day on viewing statistics. Figs. 24 and 25 plot the hourly mean play time and playback percentage over hours of a day. There is a marked increase in both metrics between 19:00 and 24:00.

This suggests that users not only watch more videos during those hours, but they also watch them for a longer time/proportion as well, likely because those hours are their time of leisure.

Video length – Intuitively, one would expect the play time to increase with the video length. Indeed, as Fig. 26 shows, the mean play time does increase with the video length.

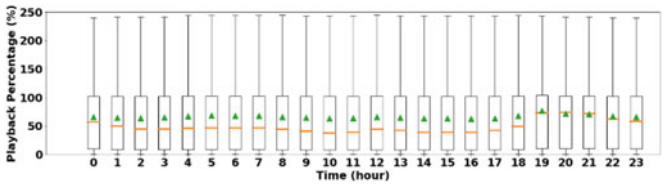

Fig. 25. Playback percentage across time-of-day.

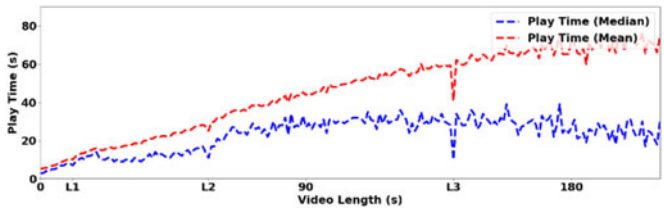

Fig. 26. Relation between play time and video length.

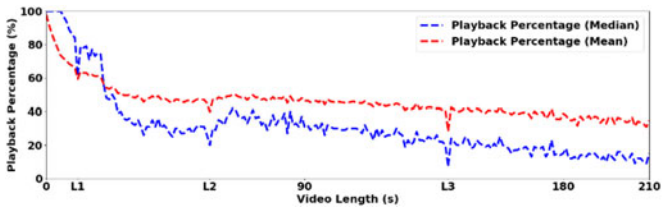

Fig. 27. Relation between playbck percentage and video length.

However, the median play time exhibits a different trajectory. Its increasing trend levels off beyond around 90 s and then begins to decrease slightly for longer video lengths (e.g., over 130 s). One possible explanation for this difference is as follows.

Short-video service users generally have very low levels of patient, i.e., they tend to terminate and switch to another video unless they find the content appealing right from the beginning. Therefore, for the less appealing contents (e.g., the bottom half), the user will only watch for so long before switching to another video. Hence, the plateau of the median playback time, i.e., around 25 s, represents the users' patient threshold.

This is significant for content producers as it means that a video must be able to attract viewer's attention within the first 25 s or else the user will likely skip. Similar observation can be drawn from the playback percentage results in Fig. 27 where the median playback percentage drops sharply for video length longer than around 25 s.

Another notable observation from Figs. 26 and 27 is the dips in both metrics at the preset video lengths L_1, L_2, and L_3. This result is consistent with the findings from the video popularity versus video length plot in Fig. 9 in Section 4.3 which shows that live video uploads are significantly less popular than pre-recorded video uploads.

We further quantify its impact in Table 8 by comparing the viewing statistics of live video uploads (with video lengths of L_1, L_2, and L_3) to those of pre-recorded video uploads (video lengths of preset lengths plus 1 s). It is evident that live video uploads were significantly less popular, more so with longer preset lengths. The mean number of accesses for the three preset durations L_1, L_2, and L_3 were merely 40%, 23%, and 8% of their pre-recorded counterparts.

TABLE 8
Playback Statistics for Live (L1, L2, L3) vs. Pre-recorded (the +1's) Video Uploads

Video Length (s)	Mean Playback Percentage	Median Playback Percentage	Mean Number of Accesses
L1	58.80%	60%	4.47
L1 + 1	63.23%	77%	11.30
L2	39.93%	20%	5.38
L2 + 1	46.84%	31%	23.37
L3	29.10%	8%	3.63
L3 + 1	43.32%	25%	44.29

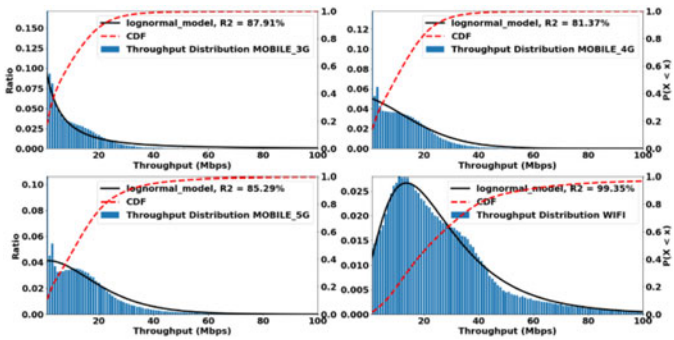

Fig. 28. Throughput across network types in dataset #3 (16^{th}-17^{th}). For mobile networks (3G, 4G, 5G), there were substantially more (over 10%) low throughput cases (less than 1Mbps) than in dataset #4, possibly due to mobile network rate-limiting.

This surprising finding could have far-reaching impact to designs in content recommendation, resource allocation, and more fundamentally, user interface design for short video services.

7 Discussions and Explorations

In this section, we consolidated some of the key findings in the previous analysis and discussed their broader implications as well as their potential applications.

7.1 Mobile Network Rate-Limiting

A significant finding from analyzing the network analytics across the four datasets is the extend of mobile network rate-limiting and its potential impact on network and service performance. The first clue arises from inconsistency in the throughput model fit across datasets.

In Section 5.2, the throughput distributions of all four network types can be approximated by lognormal models. However, we discovered some anomalies when applying it to dataset #3, shown in Fig. 28. The distributions for 4G and 5G networks in dataset #3 exhibited significantly more low-throughput cases than predicted by the model.

Further investigation suggests that this could be caused by the rate-limiting policy of the mobile operators [45]. Most operators offer data SIM with a data quota (e.g., 5 GB) for network access at either full-speed or at high-speed. Once the quota is exhausted, the operator will limit the maximum bandwidth available to the subscriber to a much lower data rate (e.g., 1 Mbps or less). This is a compromise between unlimited data (which is available but very costly) and fixed data quota (which is cheaper but very inconvenient once the quota is exhausted). Most importantly, most data SIM plans in China adopt an accounting cycle according to calendar months.

TABLE 9
Ratio of Requests with Throughput Less Than 1 Mbps Over Three Specific Days of a Month

	Day of the Month (Dataset#4)		
	02^{nd}	06^{th}	10^{th}
4G	3.72%	5.71%	9.66%
Wi-Fi	1.26%	1.51%	1.42%

TABLE 10
Rebuffering Rate Over Three Specific Days of a Month

	Day of the Month (Dataset#4)		
	02^{nd}	06^{th}	10^{th}
4G	1.37%	1.57%	2.21%
Wi-Fi	0.75%	0.92%	0.82%

In other words, the effects of rate-limiting across users are *synchronized* and will increase as one gets closer to the end of the calendar month. This explains the anomaly in dataset #3 as those were captured between 16^{th}-17^{th} of the month versus 2^{nd}-10^{th} of the month in dataset #4. Taking the analysis further, we calculated in Table 9 the ratio of low-throughput (< 1 Mbps) requests for three specific days of the month from dataset #4. We only use data from dataset #4 as it spans a wider range of days, and they were all captured within the same month, thus minimizing the impact of potential confounding factors such as improvement in network infrastructure.

The results in Table 9 show that the low-throughput ratio increases substantially over time. As a control, we calculated the same ratios for Wi-Fi which do not exhibit such an increasing trend. Another remarkable finding is how early rate-limiting kicks in – by the 10^{th}, the ratio has already increased by 2.6 times. In retrospect, perhaps this is to be expected as our samples were for short-video users where video streaming is inherently data-intensive.

Mobile network rate-limiting could have a significant impact on service performance. For example, we compare in Table 10 the rebuffering rate of those three days for 4G and Wi-Fi requests. The 4G rebuffering rate increased from 1.37% on the 2^{nd} to 2.21% on the 10^{th}, while the control, i.e., Wi-Fi rebuffering rate, showed no such trend. Worst still, if the limited rate is lower than the Service's video bitrate, the Service will become unusable over the mobile network.

The above finding could just be the tip of the iceberg as the impact of rate-limiting goes far beyond short-video services. This is uncharted territory and given the widespread deployment of rate-limiting in mobile networks, a more thorough understanding of its characteristics, impact, and detection could lead to improved designs at all levels of mobile services.

7.2 Exploring ABR for Short Video Streaming

Given that adaptive streaming has been widely deployed in many other streaming services, it begs the question of why

TABLE 11
Exploratory Experiment Settings

Parameter	Value
Video Segment Size	4 s
Initial Video Quality	750 Kbps (2nd lowest bitrate)
Bandwidth Trace Data	Pensieve Trace [46]
Video Length	8 s to 156 s
Bitrate Ladder	[300, 750, 1200, 1850, 2850, 4300] Kbps
Startup Penalty (μ_s)	4.3
Rebuffering Penalty (μ)	4.3

short-video services have yet to adopt them. We explore this question in this section by applying three well-known ABR algorithms, namely Pensieve [46], MPC [47], Buffer Based Algorithm (BBA) [48], to a short-video service streaming simulator based on Mao *et al*. [46] to explore their potential performance and limitations.

Table 11 summarizes the simulation settings. We kept the ABR algorithms' original settings except for video length, which now ranges from 8 to 156 s. We note that the original simulator [46] does not count startup delay in calculating QoE. This may not be suited for evaluating short-video services as startup delay is one of the key KPIs. Therefore, we adopted two versions of QoE metrics, $\text{QoE}_{\text{norm1}}$ and $\text{QoE}_{\text{norm2}}$:

$$\text{QoE}_{\text{norm1}} = \frac{1}{N-1}\left(\sum_{n=2}^{N} R_n - \mu \sum_{n=1}^{N} T_n - \sum_{n=1}^{N-1} |R_{n+1} - R_n|\right) \quad (6)$$

$$\text{QoE}_{\text{norm2}} = \frac{1}{N}\left(\sum_{n=1}^{N} R_n - \mu_s T_0 - \mu \sum_{n=1}^{N} T_n - \sum_{n=1}^{N-1} |R_{n+1} - R_n|\right) \quad (7)$$

where N is the total number of segments. R_n, is the video bitrate for segment n, T_n is the rebuffering duration for segment n.

$\text{QoE}_{\text{norm1}}$ measures video quality, rebuffering, and quality variation. It is the same as QoE_{lin} [46], [47], [48] adopted widely in the literature except that it is normalized by the number of video segments, as unlike in the original studies [46], the video length is variable instead of fixed. $\text{QoE}_{\text{norm2}}$ expands $\text{QoE}_{\text{norm1}}$ by incorporating the penalty due to startup delay (second term in (7)).

Our objective is to explore the impact of video length and startup delay penalty on the ABR algorithms as these are two of the key differences between current streaming services and short-video services. We first evaluate the ABR algorithms' QoE performance versus video length using $\text{QoE}_{\text{norm1}}$ in Fig. 29. In the same figure, we also plotted the

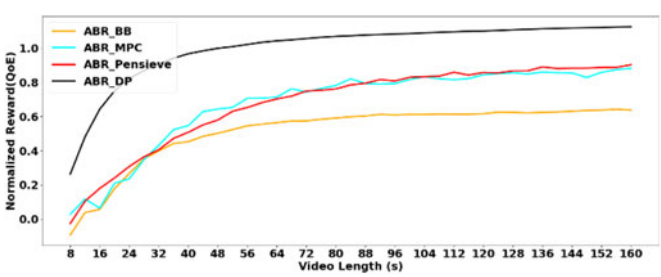

Fig. 29. The impact of video length on $\text{QoE}_{\text{norm1}}$ (w/o startup delay penalty) for Pensieve, Robust MPC, BBA, and Offline optimal (DP).

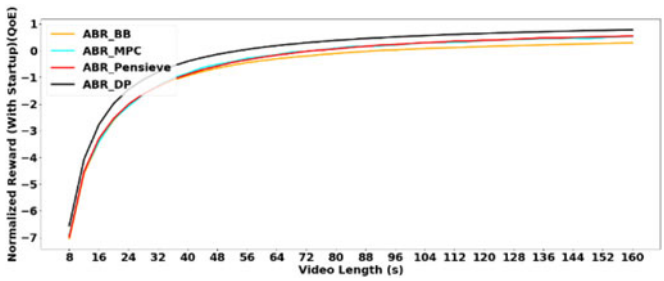

Fig. 30. The impact of video length on $\text{QoE}_{\text{norm2}}$ (w/ startup delay penalty) for Pensieve, Robust MPC, BBA, and offline optimal (DP).

offline optimal QoE (assuming full knowledge of future bandwidth) obtained from dynamic programming [46]. There are two important observations.

First, it is clear that video length has a significant impact on the QoE performances of all three ABR algorithms. Specifically, their QoE decreases as video length shortens. The degradation accelerates when video length falls below around 40 s. Note that we assumed videos are always played completely in the simulator, so the video length is more akin to video play time in practice. In the Service, only 23.35% of streaming sessions have play time longer than 40 s, implying that current ABR algorithms would likely operate in the degraded performance regime most of the time.

Second, even the offline optimal exhibited similar performance degradations at short video lengths. One reason for the degradation is the initial bitrate choice for the first video segment, which defaults to the second bitrate version (750 Kbps). While this choice may be inconsequential in conventional streaming services with longer video lengths, its impact on short video streaming increases rapidly as video length shortens. For example, with a 20-s video, the initial segment (4 s) already accounts for 20% of the whole video. While the ABR algorithms will subsequently raise the bitrate when bandwidth allows, climbing the bitrate ladder also results in bitrate variation penalties (i.e., the third term in (6)) which offset the gains in video quality. In conventional streaming, the ABR algorithm can still reap the benefits once the bitrate converges as the video length is typically longer. By contrast, in short-video streaming, the ABR algorithm may not even have sufficient time to converge, let alone reap the benefits from an optimized bitrate choice.

Next, we explore the impact of startup-delay penalty in Fig. 30, which plots $\text{QoE}_{\text{norm2}}$ versus video length for the four schemes. The QoE performances degraded even further by the startup-delay penalty, so much so that they turn negative for a video shorter than 56 s. Obviously, the results depend on the choice of the weighting coefficient for the startup-delay penalty – it was set to be the same as the one for the rebuffering penalty (Table 11) in this exploratory study.

The results from the above exploratory experiment are obviously not meant to be conclusive, but rather to demonstrate the potential performance limitations of current ABR algorithms when applied to short-video streaming. The surprisingly large performance impacts of short video length and startup delay penalty call for new thinking not only in the design of ABR algorithms, but also in the formulation of new QoE functions for short-video services.

8 Summary and Future Work

The extensive data analysis in this measurement study offers new insights into the characteristics and behavior of a large-scale short-video service. The measurement results, many of which could be modeled by well-known mathematical distributions, could be readily employed in the study of short-video services, either as inputs to improve the fidelity of simulators, or as the basis to formulate a system model for performance analysis and optimization. In addition to the models presented in the main text, readers can also find additional models in the appendices.

Furthermore, the measurement results revealed many short-video service characteristics that differ significantly from conventional streaming services. Examples include the extremely rapid popularity evolution; the conformance to the Zipf law at a very short timescale; strong correlations between popularity, upload time, and video length; the surprising popularity difference between pre-recorded versus live video uploads; the correlation between connection time and throughput; the impact of mobile network rate-limiting; the very short user-patient; and so on.

Many of these findings point to new ways to design and optimize the content delivery for short-video services, which are important topics for future research given the reach and scale of such services grown over just a few years. Moreover, the exploratory experiment in Section 7.2 clearly demonstrated the limitations of applying current ABR algorithms to short-video services, which calls for new thinking not only in the design, but also the in evaluation of these services.

Acknowledgments

The authors would like to thank the associate editor and the anonymous reviewers for their insightful comments and suggestions in improving this paper and the generous support of the anonymous short video service operator for providing the anonymized application log data that enabled this study.

References

[1] Data story of Tik-Tok. Accessed: Jan. 01, 2021. [Online]. Available: https://www.politesi.polimi.it/handle/10589/152682

[2] Baidu document. Accessed: Jan. 01, 2021. [Online]. Available: https://wenku.baidu.com/view/fa8cdcbb5cf7ba0d4a7302768e9951e79b896922.html

[3] Sina tech. Accessed: Jan. 01, 2021. [Online]. Avaliable: https://tech.sina.com.cn/csj/2020-02-22/doc-iimxxstf3533730.shtml

[4] M. Zink, K. Suh, Y. Gu, and J. Kurose, "Characteristics of youtube network traffic at a campus network - Measurements, models, and implications," *Comput. Netw.*, vol. 53, no. 4, pp. 501–514, Mar. 2009.

[5] X. Cheng, J. Liu, and C. Dale, "Understanding the characteristics of internet short video sharing: A youtube-Based measure-ment study," *IEEE Trans. Multimedia*, vol. 15, no. 5, pp. 1184–1194, Aug. 2013.

[6] P. Gill, M. Arlitt, Z. Li, and A. Mahanti, "youTube traffic characterization: A view from the edge," in *Proc. ACM SIGCOMM Conf. Internet Meas.*, 2007, pp. 15–28.

[7] M. Ghasemi, P. Kanuparthy, A. Mansy, T. Benson, and J. Rexford, "Performance characterization of a commercial video streaming service," in *Proc. ACM SIGCOMM Conf. Internet Measu.*, 2016, pp. 499–511.

[8] M. Plakia *et al.*, "Should i stay or should i go: Analysis of the impact of application QoS on user engagement in youtube," *ACM Trans. Model. Perform. Eval. Comput. Syst.*, vol. 5, no. 2, pp. 1–32, Apr. 2020.

[9] M. Cha, H. Kwak, P. Rodriguez, Y. Ahn, and S. Moon, "I tube, you tube, everybody tube: Analyzing the world's largest user generated content video system," in *Proc. ACM SIGCOMM Conf. Internet Meas.*, 2007, pp. 1–14.

[10] L. Plissonneau and E. Biersack, "A longitudinal view of HTTP video streaming performance," in *Proc. 3rd Multimedia Syst. Conf.*, 2012, pp. 203–214.

[11] P. Casas, P. Fiadino, A. Sackl, and A. D'Alconzo, "YouTube in the move: Understanding the performance of youtube in cellular networks," in *Proc. IFIP Wireless Days*, 2014, pp. 1–6.

[12] A. Finamore, M. Mellia, M. Munafò, R. Torres, and S. G. Rao, "YouTube everywhere: Impact of device and infrastructure synergies on user experience," in *Proc. ACM SIGCOMM Conf. Internet Meas.*, 2011, pp. 345–360.

[13] Y. Chen *et al.*, "Understanding viewer engagement of video service in wi-fi network," *Comput. Netw.*, vol. 91, no. 14, pp. 101–116, Nov. 2015.

[14] M. Shafiq, J. Erman, L. Ji, A. Liu, and J. Pang, "Understanding the impact of network dynamics on mobile video user engagement," in *Proc. SIGMETRICS Int. Conf. Meas. Model. Comput. Syst.*, 2014, pp. 16–20.

[15] F. Figueiredo, F. Benevenuto, and J. Almeida, "The tube over time: Characterizing popularity growth of youtube videos," in *Proc. ACM Int. Conf. Web Search Data Mining*, 2011, pp. 745–754.

[16] A. Jia, S. Shen, D. Li, and S. Chen, "Predicting the implicit and the explicit video popularity in a user generated content site with enhanced social features," *Comput. Netw.*, vol. 140, no. 20, pp. 112–125, Jul. 2018.

[17] A. Jia, S. Shen, S. Chen, D. Li, and A. Iosup, "An analysis on a youtube-like UGC site with enhanced social features," in *Proc. Int. Conf. World Wiide Web Companion*, 2017, pp. 1477–1483.

[18] Z. Chen, Q. He, Z. Mao, and H. M. Chung, "A study on the charac-teristics of douyin short videos and implications for edge caching," in *Proc. ACM Turing Celebration Conf.*, 2019, pp. 1–6.

[19] J. He, M. Hu, Y. Zhou, and D. Wu, "LiveClip: Towards intelligent mobile short-form video streaming with deep reinforcement learning," in *Proc. 30th ACM Workshop Netw. Oper. Syst. Support Digit. Audio Video*, 2020, pp. 54–59.

[20] L. Zhang, F. Wang, and J. Liu, "Mobile instant video clip sharing with screen scrolling: Measurement and enhancement," *IEEE Trans. Multimedia*, vol. 20, no. 8, pp. 2022–2034, Aug. 2018.

[21] L. Cherkasova and M. Gupta, "Characterizing locality, evolution, and life span of accesses in enterprise media server workloads," in *Proc. 12th Int. Workshop Netw. Oper. Syst. Support Digit. Audio Video*, 2002, pp. 33–42.

[22] W. Tang, Y. Fu, L. Cherkasova, and A. Vahdat, "Modeling and generating realistic streaming media server workloads," *Comput. Netw.*, vol. 51, no. 1, pp. 336–356, Jan. 2007.

[23] Y. Zhang, P. Li, Z. Zhang, B. Bai, and G. Zhang, "AutoSight: Distributed edge caching in short video network," *IEEE Netw.*, vol. 34, no. 3, pp. 194–199, May/Jun. 2020.

[24] Y. Zhang *et al.*, "Challenges and chances for the emerging short video network," in *Proc. IEEE Conf. Comput. Commun. Workshops*, 2019, pp. 1025–1026.

[25] S. S. Krishnan and R. K. Sitaraman, "Video stream quality impacts viewer behavior: Inferring causality using quasi-experimental designs," *IEEE Trans. Netw.*, vol. 21, no. 6, pp. 2001–2014, Dec. 2013.

[26] F. Dobrian *et al.*, "Understanding the impact of video quality on user engagement," in *Proc. ACM SIGCOMM Conf.*, 2011.

[27] Y. Ding, Y. Du, Y. Hu, Z. Liu, L. Wang, and K. Ross, "Broadcast yourself: Understanding youtube uploaders," in *Proc. ACM SIGCOMM Conf. Internet Meas.*, 2011, pp. 2–4.

[28] J. Jiang, R. Das, and G. Ananthanarayanan, "Via: Improving internet telephony call quality using predictive relay selection," in *Proc. ACM SIGCOMM Conf.*, 2016, pp. 286–299.

[29] M. Arantes, F. Figueiredo, and J. M. Almeida, "Understanding video-ad consumption on youtube: A measurement study on user behavior, popularity, and content properties," in *Proc. Conf. Web Sci.*, 2016, pp. 25–34.

[30] A. Zhou, H. Zhang, G. Su, L. Wu, R. Ma, and Z. Meng, "Learning to coordinate video codec with transport protocol for mobile video telephony," in *Proc. Annu. Int. Conf. Mobile Comput. Netw.*, 2019, pp. 1–16.

[31] T. Stockhammer, "Dynamic adaptive streaming over HTTP: Standards and design principles," in *Proc. ACM Multimedia Syst.*, 2011, pp. 133–144.

[32] *Network Protocols Handbook*, 2nd ed. Javvin Technologies Inc., Saratoga, CA, USA, 2005.
[33] K. M. Charles, *The TCP/IP Guide: A Comprehensive, Illustrated Internet Protocols Reference*, 1st ed. San Francisco, CA, USA: No Starch Press, 2005.
[34] A. Ferragut, I. Rodríguez, and F. Paganini, "Optimizing TTL caches under heavy-tailed demands," in *Proc. Int. Conf. SIGMETRICS Meas. Model. Comput. Syst.*, 2016, pp. 14–18.
[35] G. Zipf, *Human Behavior and the Principle of Least Effort*. Boston, MA, USA: Addison-Wesley, 1949.
[36] R. G. D. T. Steel and H. J., *Principles and Procedures of Statistics with Special Reference to the Biological Sciences*. New York, NY, USA: McGraw Hill, 1960.
[37] J. Wang, Y. Zheng, Y. Ni, C. Xu, F. Qian, and W. Li, "An Active-passive measurement study of TCP performance over LTE on high-speed rails," *IEEE Commun. Surv. Tuts.*, vol. 19, no. 3, pp. 1842–1866, 2017.
[38] A. Langley, A. Riddoch, A. Wilk, and A. Vicente, "The QUIC transport protocol: Design and internet-scale deployment," in *Proc. Ann. Conf. ACM Special Int. Group Data Commun.*, 2017, pp. 21–25.
[39] S. Li, J. Xu, and M. Van Der Schaar, "Popularity-driven content caching," in *Proc. Int. Conf. Comput. Commun.*, 2016, pp. 10–14.
[40] *5G Handbook,* ShareTechnote. Accessed: Jan. 01, 2021. [Online]. Avaiable: https://www.sharetechnote.com/html/5G/Handbook_5G_Index.html
[41] H. S. Goian, O. Y. Al-Jarrah, and S. Muhaidat, "Popularity-based video caching techniques for cache-enabled networks: A survey," *IEEE Access*, vol. 7, pp. 27699–27719, 2019.
[42] T. Syeda-Mahmood and D. Ponceleon, "Learning video browsing behavior and its application in the generation of video previews," in *Proc. 9th ACM Int. Conf. Multimedia*, 2011, pp. 119–128.
[43] L. Chen, Y. Zhou, and D. M. Chiu, "Video browsing-A study of user behavior in online VoD services," in *Proc. Int. Conf. Comput. Commun. Netw.*, 2013, pp. 1–7.
[44] L. Guo and J. Y. B. Lee, "Stateful-TCP—A new approach to accelerate TCP slow-start," *IEEE Access*, vol. 8, pp. 195 955–195 970, 2020.
[45] T. Flach, P. Papageorge, A. Terzis, and L. Pedrosa, "An internet-wide analysis of traffic policing," in *Proc. ACM SIGCOMM Conf.*, 2016, pp. 468–482.
[46] H. Mao, R. Netravali, and M. Alizadeh, "Neural adaptive video streaming with pensieve," in *Proc. Conf. ACM Special Int. Group Data Commun.*, 2017, pp. 197–210.
[47] X. Yin, A. Jindal, V. Sekar, and B. Sinopoli, "A control-theoretic approach for dynamic adaptive video streaming over HTTP," in *Proc. Conf. ACM Special Int. Group Data Commun.*, 2015, pp. 325–338.
[48] T. Y. Huang, R. Johari, N. McKeown, and M. Trunnell, "A buffer-based approach to rate adaptation: Evidence from a large video streaming service," in *Proc. Conf. SIGCOMM*, 2014, pp. 187–198.

Yuming Zhang received the BEng degree in information engineering from the Chinese University of Hong Kong, Shatin, Hong Kong, in 2018. He is currently working toward the PhD degree with the Department of Information Engineering, Chinese University of Hong Kong. His research interests include adaptive video steaming, caching, and protocol optimization.

Yan Liu received the BEng degree in communication engineering from the University of Electronic Science and Technology of China Chengdu, China, in 2012 and the PhD Degree in information engineering from the Chinese University of Hong Kong, Shatin, Hong Kong, in 2016. He is currently a senior engineer with the Cloud ARCH and Platform Department, Tencent, China. His research interests include computer networks, including but not limited to Internet congestion control and video streaming.

Lingfeng Guo received the BEng in software engineering from Sun Yat-Sen University, Guangdong, China, in 2016 and the PhD degree in information engineering from the Chinese University of Hong Kong, Shatin, Hong Kong, in 2020. He is currently a senior engineer with the Cloud ARCH and Platform Department, Tencent, China, where he participates in the research and development of Internet protocols.

Jack Y. B. Lee (Senior Member, IEEE) received the BEng and PhD degrees in information engineering from the Chinese University of Hong Kong, Shatin, Hong Kong, in 1993 and 1997, respectively. He is currently an associate professor with the Department of Information, Chinese University of Hong Kong. His research interests include research in multimedia communications systems, mobile communications, protocols, and applications. He specializes in tackling research challenges arising from real-world systems. He works closely with the industry to uncover new research challenges and opportunities for new services and applications. Several of the systems research from his lab have been adopted and deployed by the industry.