Understanding Performance of Edge Content Caching for Mobile Video Streaming

Ge Ma, Zhi Wang, *Member, IEEE*, Miao Zhang, Jiahui Ye, Minghua Chen, *Senior Member, IEEE*, and Wenwu Zhu, *Fellow, IEEE*

***Abstract*— Today's Internet has witnessed an increase in the popularity of mobile video streaming, which is expected to exceed 3/4 of the global mobile data traffic by 2019. To satisfy the considerable amount of mobile video requests, video service providers have been pushing their content delivery infrastructure to edge networks—from regional content delivery network (CDN) servers to peer CDN servers (e.g., smartrouters in users' homes)—to cache content and serve users with storage and network resources nearby. Among the edge network content caching paradigms, Wi-Fi access point caching and cellular base station caching have become two mainstream solutions. Thus, understanding the effectiveness and performance of these solutions for large-scale mobile video delivery is important. However, the characteristics and request patterns of mobile video streaming are unclear in practical wireless network. In this paper, we use real-world data sets containing 50 million trace items of nearly 2 million users viewing more than 0.3 million unique videos using mobile devices in a metropolis in China over two weeks, not only to understand the request patterns and user behaviors in mobile video streaming, but also to evaluate the effectiveness of Wi-Fi and cellular-based edge content caching solutions. To understand the performance of edge content caching for mobile video streaming, we first present *temporal* and *spatial* video request patterns, and we analyze their impacts on caching performance using frequency-domain and entropy analysis approaches. We then study the behaviors of mobile video users, including their mobility and geographical migration behaviors, which determine the request patterns. Using trace-driven experiments, we compare strategies for edge content caching, including least recently used (LRU) and least frequently used (LFU), in terms of supporting mobile video requests. We reveal that content, location, and mobility factors all affect edge content caching performance. Moreover, we design an efficient caching strategy based on the measurement insights and experimentally evaluate its performance. The results show that our design significantly improves the cache hit rate by up to 30% compared with LRU/LFU.**

***Index Terms*— Edge network, mobile video streaming, user behavior, measurement, content delivery.**

I. Introduction

GLOBAL mobile video traffic reached 3.7 EB/month at the end of 2015, and it is predicted that over 3/4 of the global mobile data traffic will be video traffic by 2019 [1]. This trend is further accelerated by the rapid growth of online/mobile social media and mobile networks: video clips are increasingly being generated by users and instantly shared with their friends. In contrast to conventional live and on-demand video streaming that are consumed using TVs and PCs, mobile video streaming is generally watched by users on mobile devices with wireless connections, i.e., 3G/4G cellular or Wi-Fi. User behaviors and wireless network quality in mobile video streaming [2], [3] can be quite different from those in conventional video streaming [4], [5], thus requiring improvements in the delivery of mobile video streaming.

To meet the sky-rocketing increase in bandwidth requirements resulting from data-intensive video streaming and to reduce the monetary cost for renting expensive resources in conventional content delivery networks (CDNs), video service providers are pushing their content delivery infrastructure closer to users to utilize network and storage resources in households for content delivery [6], including caching content over femtocells [7] and replicating video content via Wi-Fi smartrouters in households. Youku, one of the largest online video providers in China, has deployed over 300K smartrouters in its usersâŁ™ homes in less than one year, expecting to transform a large fraction of its users (250M) into such content delivery peer nodes [8]. To serve users with good quality of experience using the new edge network solutions, it is important to answer the following questions: (1) What are the video request patterns in mobile video streaming, how do users behave in today's mobile video systems, and what is the implication of their behaviors on edge network video content delivery (Sec. IV and Sec. V)? (2) How is the quality of user experience in the mobile video sessions (Sec. VI-B)? (3) Can today's mobile network infrastructure appropriately satisfy the mobile video streaming demand (Sec. VI-A)?

Manuscript received September 22, 2016; revised January 13, 2017; accepted January 26, 2017. Date of publication March 9, 2017; date of current version May 24, 2017. This work was supported in part by the National Natural Science Foundation of China under Grant U1611461, Grant 61402247 and Grant 61531006, in part by the National Basic Research Program of China (973) under Grant 2015CB352300 and Grant 2013CB336700, in part by the University Grants Committee of the Hong Kong Special Administrative Region, China, under Grant C7036-15G, and in part by the research fund of Tsinghua-Tencent Joint Laboratory for Internet Innovation Technology.

G. Ma and J. Ye are with the Tsinghua-Berkeley Shenzhen Institue, Tsinghua University, Shenzhen 518055, China (e-mail: mg15@mails.tsinghua.edu.cn; yejh16@mails.tsinghua.edu.cn).

Z. Wang and M. Zhang are with the Graduate School at Shenzhen, Tsinghua University, Shenzhen 518055, China (e-mail: wangzhi@sz.tsinghua.edu.cn; zhangmia15@mails.tsinghua.edu.cn).

M. Chen is with the Department of Information Engineering, The Chinese University of Hong Kong, Hong Kong (e-mail: minghua@ie.cuhk.edu.hk).

W. Zhu is with the Tsinghua National Laboratory for Information Science and Technology, Beijing Key Laboratory of Networked Multimedia, Department of Computer Science and Technology, Tsinghua-Berkeley Shenzhen Institue, Tsinghua University, Beijing 100084, China (e-mail: wwzhu@tsinghua.edu.cn).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/JSAC.2017.2680958

(4) What strategies can be applied to best support mobile video content delivery (Sec. VII)?

Several measurement studies have been conducted to address the above questions. However, such measurement studies are challenging because many different factors are involved, including user behaviors (i.e., mobility pattern and video preference), video content characteristics, and mobile network characteristics. Previous studies generally focus on a single aspect, e.g., studying the popularity of mobile video content [9], [10], user mobility behaviors [11], or network strategies to support mobile video streaming, e.g., content replication [12]. The limitation of the previous studies is that they have not considered the joint impact of user behaviors, content characteristics and wireless network deployment, on edge network content delivery.

In this paper, we propose to address the above questions from the perspectives of both the mobile video service and wireless network providers. From the perspective of mobile video service, we study how users view mobile videos, including their mobility patterns in video sessions and the content selection in different locations, and we build a mobile video consumption model. From the perspective of wireless network provider, we present how the mobile video requests can be served by both the Wi-Fi and cellular infrastructures that are commonly used by today's users, and we provide insights on how to improve the QoS of wireless networks according to their video request patterns.

Our contributions are summarized as follows:

First, we use large-scale datasets to study user behaviors in a real-world mobile video streaming system, covering 50 million sessions of nearly 2 million users viewing more than 0.3 million videos using mobile devices in 2 weeks. Using frequency-domain and entropy analyses [9], [13], we show that mobile video requests exhibit unique spatial and temporal patterns that can significantly affect the performance of content caching strategies in edge networks. (1) We observe a skewed geographic request distribution in the mobile video system, and the number of requests is highly affected by the regular mobility patterns of users. For example, the number of requests in train stations is much larger than that in residential areas. (2) We observe that videos with lower popularity have more uniform distribution of requesting locations, while videos with higher popularity have more skewed distribution. Surprisingly, the increase of multi-location users (who request videos in different locations, i.e., mobile users) in a location does not increase the requested number of unique videos, which is different from conventional single-location users (who request videos only in the same locations, i.e., home users). (3) In the frequency-domain analysis, we observe that the number of requests in locations with different functionalities over time has 3 major periods, e.g., residential areas have an obvious period of 8 hours, which can be used to predict the future traffic in content delivery.

Second, we further investigate how user behaviors determine the above request patterns. We reveal that both mobility and geographic migration behaviors of users can significantly affect mobile video requests. In particular, we show that the mobility behaviors of users are heterogeneous, e.g., a number of multi-location users request videos intensively and request them in different locations, whereas there is a large fraction of users who only request a small number of videos in the same location. For the geographic migration behaviors, we observe that (1) users have regular commute behaviors, involving 2–3 regularly visited locations where they tend to request mobile videos, and (2) it is common for users to move between the same type of locations (e.g., residential) and issue video requests. *These observations suggest that joint caching strategies over multiple locations can improve the caching performance*.

Third, we compare the effectiveness of Wi-Fi and cellular-based edge network caching solutions, and we discuss the potential improvement on mobile video streaming to today's wireless networks. Based on our edge network traces covering $1,055,881$ Wi-Fi APs and $69,210$ cellular base stations, we investigate conventional caching strategies, including least recently used (LRU) and least frequently used (LFU) for edge network mobile video delivery. We first show that most of today's Wi-Fi and cellular deployments are close enough to the mobile requests of users in different locations; however, although Wi-Fi and cellular have different deployment strategies, they cannot well serve different categories of mobile video users. Second, we show that a number of factors including user mobility, content popularity, cache capacity, and caching strategies affect the caching performance for both Wi-Fi and cellular caching for mobile video delivery. For example, unpopular videos attract users mostly from few locations where users have particular interests in the content, and caching strategies have various influences on different categories of users.

Finally, motivated by the measurement insights, we design a geo-collaborative caching strategy for mobile video delivery, which jointly considers mobile video request patterns, user behaviors and the deployment of wireless networks. Based on real-world trace-driven experiments, we show that our design achieves a 20% (resp. 30%) cache hit rate improvement and a 20% (resp. 30%) service rate improvement compared with conventional LRU (resp. LFU) caching strategies.

The remainder of this paper is organized as follows. We discuss the related works in Sec. II. We present the datasets used in our measurement studies in Sec. III. We study the temporal and spatial request patterns and the content characteristics in a mobile video system in Sec. IV. We measure user behaviors in mobile video streaming sessions and how they affect the quality of mobile streaming in Sec. V. We compare the effectiveness of Wi-Fi and cellular-based edge content delivery solutions and discuss the potential improvement to today's wireless networks to improve mobile video streaming in Sec. VI. We present the details of our caching strategy and evaluate its performance in Sec. VII. Finally, we conclude the paper in Sec. VIII.

II. Related Work

There are four main research areas related to our work: video measurement, user mobility behaviors, edge video delivery and edge network caching strategies.

A. Video Measurement

There are several prior studies that focus on the properties of videos and how to model and predict the popularity of such videos. One of the works investigates the relationship between the popularity and location of online videos [9], [14]. This work finds that videos exhibit a geographical distribution of interest, with users arising from a confined and single area rather than from a global area, and it provides new insights on how the geographic reach of a video changes as its popularity peak and then fades away. The prediction of video popularity has also been studied based on historical information given by early popularity measures [10], [15]. Two novel models are proposed, which are able to better distinguish between videos with different popularities, by assigning different weights to samples with different popularities and exploring the similarity between the video and known samples within the monitoring period. Our study on mobile video differs from these works since our analysis focuses not only on time period (hour level), but also on entropy analysis. In addition, the geographic locations that we measured are more specific, allowing us to obtain a comprehensive relationship between temporal and spatial video request patterns.

B. User Mobility Behaviors

There are also several prior studies that focus on characterizing mobile video traffic. Li *et al.* [16] focus on analyzing the main discrepancies when users access video-on-demand systems using either Wi-Fi or 3G connections. They study the factors that affect mobile users' interests and video popularity. Li *et al.* [14] characterize the geographical patterns on a large-scale, commercial, mobile video-on-demand system and analyze the temporal evolution trends of the geographical popularity, which reveal distinct behaviors of popular and unpopular videos. However, they only use coarse-grained (in province) location information, which differs from our study in which the latitudes and longitudes of users are analyzed to obtain useful insights about the relationship between user mobility and video request patterns. Recently, Wang *et al.* [13] model the mobile traffic patterns of large-scale cellular base stations deployed in a city. Their work contributes some valuable information for Internet service providers, mobile users, and government management of mobile network resources.

With the development of new location-sensing technologies, the information about the locations of users has become available. Toole *et al.* [17] use the dynamic data generated by mobile phones to measure spatiotemporal changes in population, and identify the relationship between land use and dynamic population. Considering that sharing precise location information may cause leaks of privacy information, Das *et al.* [11] study the contextual locations of users by passively monitoring the mobile network traffic of many location-based services, which only rely on contextual location. In contrast to these works, our study focuses on providing an understanding of user mobility and geographical migrations when using mobile video services. A QoE modeling framework with user, system and context components is created for a mobile video environment, taking mobile user, mobile device, mobile network and mobile video service into consideration [18]. Thus, users requesting mobile videos may benefit from the model, and video providers could also develop effective strategies to improve the user experience. Furthermore, the viewing conditions of mobile video can be described in terms of three main factors: display size and viewing distance, surrounding luminance, and body movements of the viewer [19]. It incorporates all three of these important factors into optimizing video coding and delivery for mobile devices. Some studies show that users' cooperation can effectively reduce the servers' burden, such as delay and bandwidth, confirmed to be an attractive solution to limit the costs incurred by content providers [20], [21].

C. Edge Video Delivery

The substantial demand for bandwidth from data-intensive applications has challenged the traditional content delivery paradigms: the content delivery network (CDN), including its variations ISP-operated CDN [22], content-provider-operated CDN [23], and peer-to-peer CDN [24], [25]. Because mobile video content has occupied most of the mobile network traffic, caching videos in the network edge (i.e., femtocells or Wi-Fi APs) has become a common solution. Building caches at the network edge is an appealing solution since the cost of network equipment, such as cellular base stations, substantially exceeds the cost of installing a cache [26]. Furthermore, if videos can be fetched from a local cache rather than CDN servers, the large delays can be significantly reduced [27]. Golrezaei *et al.* [7] envision femtocell-like base stations called helpers, with weak backhaul links but large storage capacity, which can assist in the macro base station by handling requests for popular files that have been cached. Based on a real measurement study of mobile video viewing logs from a leading Internet video provider for 14 days, Lin *et al.* [28] study the potential of peer-assisted video delivery in Wi-Fi mobile networks aiming to reduce server load. Moreover, Zhou *et al.* [29] study how video popularity changes over time and varies among different categories, and they apply the results to design video caching strategies in CDN servers.

D. Edge Network Caching Strategies

The impact of content popularity dynamics on cache performance can be captured by an analytical model under the assumption that requests at different caches are independent [30]. Based on this assumption, a threshold-based caching scheme is proposed for wireless access networks, which replicates content that is requested more times than the given threshold [31]. To investigate collaborative caching, coded caching strategies for heterogeneous wireless networks have been proposed to balance the cost among base station transmission, access point storage and user connection latency [32], [33]. Distributed caching architectures have also been proposed to replicate content close to users to reduce the average video delivery delay [34].

To the best of our knowledge, we are the first to jointly measure both the mobile video request patterns, user mobility behaviors, and the deployment of wireless networks to

TABLE I
MOBILE VIDEO BEHAVIOR DATASET

Field	Description
User ID (anonymized)	The unique identifier of each user
Request time	The specific time that the user requests a video
Latitude and longitude	The position of current request is issued
Video content	The name and some basic information of the video

TABLE II
WI-FI AND CELLULAR NETWORK DATASET

Field	Description
MAC	The MAC address of the device
Latitude and longitude	The specific position of the device
LAC and Cell IDs	The location code and cell ID of the base station
MNC ID	The mobile network code
Address of AP	The detailed address of the device
PoI	A functionality description of the location, e.g., *university*

investigate the performance of wireless network caching and to design an efficient caching strategy for mobile video delivery.

III. DATASETS ON MOBILE VIDEO STREAMING AND EDGE NETWORKS

In this section, we present how we collect the datasets used in our study.

A. Mobile Video Behavior Dataset

The mobile video behavior dataset is collected by a video provider company in China. How users view videos in the mobile video streaming app has been recorded. The dataset spans 2 weeks and covers 2 million users watching 0.3 million unique videos in Beijing. In each trace item, the following information is recorded: (1) The device identifier, which is unique for different devices and can be used to track users; (2) The timestamp when the user starts to watch the video; (3) The location where the user watches the video: the video player reports the location either collected from the device's built-in GPS function or inferred from the network parameters (e.g., cellular base station); (4) The title of the video, as shown in Table I.

B. Wi-Fi and Cellular Network Dataset

We also study how today's edge network content delivery paradigms can be supported by both Wi-Fi and cellular solutions [35], [36].

1) Wi-Fi AP Information: The Wi-Fi and cellular network dataset is provided by Tencent Wi-Fi [37], a mobile app that asks users to respond to questions on how they use Wi-Fi/cellular networks. In particular, we collected over 1 million Wi-Fi APs in Beijing city, including the basic service set identifier (BSSID) of Wi-Fi APs and the location of the Wi-Fi hotspots. This valuable dataset samples a large fraction of Wi-Fi APs that are actually deployed in Beijing, allowing us to determine whether these APs can provide content delivery functionality for mobile video streaming. Table II shows the details of the dataset: each trace item contains the latitude and longitude of the AP and the point of interest (PoI) information of the AP (e.g., hotel).

2) Cellular Base Station Information: Our dataset also contains cellular network information, including locations, IDs, and location area code (LAC) of over 70 thousand cellular base stations.

IV. REQUEST PATTERNS IN MOBILE VIDEO STREAMING

In this section, we first investigate the popularity distribution of mobile videos; we then study the spatial and temporal patterns of users' video requests in mobile video streaming and present how content affects mobile video requests; finally, we present the implications of such request patterns.

A. Popularity Distribution

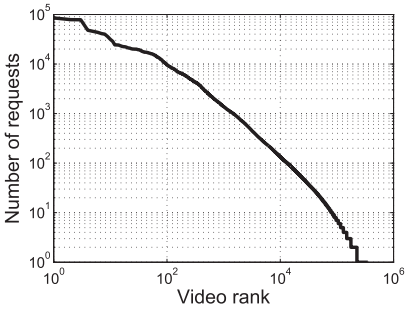
We first describe the popularity distribution of mobile videos. As illustrated in Fig. 1(a), we observe that the popularity of mobile video content also follows a power-law distribution.

Fig. 1(b) shows the average normalized number of daily requests for different video categories over time, for the 1000 most popular videos. We observe that trailer has the smallest decreasing rate, short variety show has the largest decreasing rate, and the animation category has the longest lifetime.

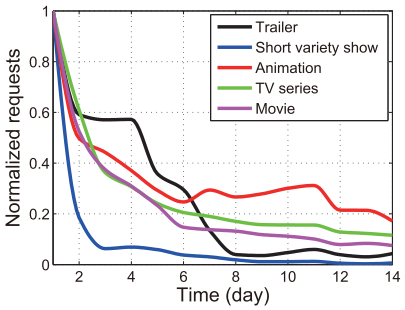
Furthermore, we investigate the popularity of videos in different locations by studying the popularity rank of the $1,000$ most popular videos (top 0.3% in the entire system). In Fig. 1(c), we plot the CDF of the average popularity rank of these videos in $1,000$ locations where they are requested. We observe that the top 0.3% videos have quite different popularity ranks in different locations: the average popularity rank for these videos is below the top 40% in as many as 60% of the locations. *This observation indicates that global popularity cannot be directly used to infer the local popularity of mobile video content.* Thus a local caching strategy is more suitable than a global strategy in current mobile video systems.

B. Spatial and Temporal Patterns of Mobile Video Requests

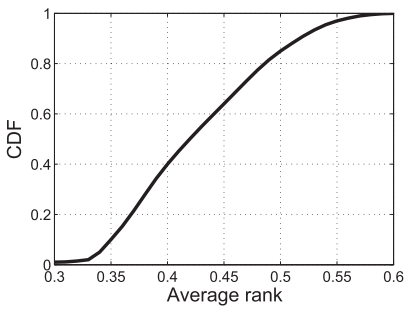
To study the mobility patterns of viewers, we assume that the users' requests can be served by the nearest Wi-Fi APs or cellular BSes. Thus, we first classify all the users in the mobile video streaming system into two categories: *multi-location users*, who request videos in different locations (APs/BSes) within *one day* in the traces, and *single-location users*, whose requests are all issued from the same location (APs/BSes) within *one day*. Note that a user may be a multi-location user or a single-location user on different days.



(a) Popularity distribution



(b) Popularity change over time



(c) Local popularity of the most popular videos in the whole system

Fig. 1. Characteristics of mobile video popularity.

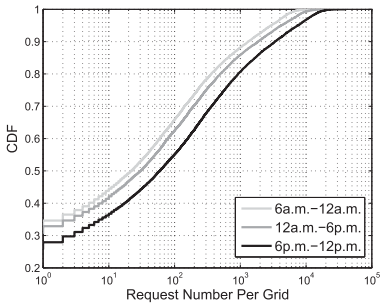


Fig. 2. CDF of the number of requests in the locations partitioned geographically.

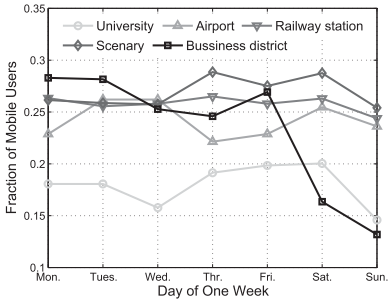


Fig. 3. Evolution of multi-location users in different locations.

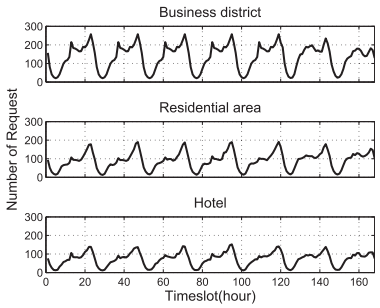


Fig. 4. Number of mobile video requests in different locations over time.

1) Skewed Geographical Request Distribution: We investigate the geographical distribution of requests. According to the longitude interval 0.01° and latitude interval 0.01°, Beijing will be divided into different locations. Every location can be abstracted as a 0.01° × 0.01° geographic location with an area of $0.72km^2$. Each location has a PoI functionality label, which indicates the largest PoI functionality number of the location. We count the number of requests issued in these locations. As shown in Fig. 2, we plot the CDFs of the number of requests issued in the locations at different times of a day, i.e., 6am–12am, 12am–6pm, and 6pm–12pm. Our observations are as follows: (1) More requests are issued at night than during the daytime, e.g., the number of requests from 6pm–12pm is 74% greater than that from 12am–6pm. (2) A significant fraction of locations only have very few requests issued. These observations indicate that to serve mobile video requests, the edge network content delivery systems need to take the geographical request distribution into consideration, e.g., to allocate more resources to the locations with higher request density and proactively push content to the edge at the off-peak times.

2) Multi-location Users in Different Locations: We study the behaviors of multi-location users in different locations within one day, such as university, airport, railway station, scenery spot and business district. In Fig. 3, we plot the fraction of multi-location users over all video users recorded in our traces in these locations in one week. Our observations are as follows: (1) These locations generally have a relatively stable multi-location user fraction of approximately 20%. (2) Some locations have lower multi-location user fractions than others, e.g., there are less users in university than at rail station. (3) *The fraction of multi-location users changes significantly over time in some locations*, e.g., the fraction in the business distinct drops from approximately 25% on weekdays to 15% on weekends. The reason is that the mobile video behaviors are highly correlated with the regular commute behaviors of users.

3) Frequency Analysis of Periodical Request Patterns: It is common for users to generate periodical requests, e.g., more video requests are issued at night. Such periodical request patterns can affect the edge network caching strategies, including content replication and resource allocation. As illustrated in Fig. 4, the curves represent the number of video requests issued in different locations in one week. The requests over time have different periodical patterns.

To specify the periodical request patterns, we use a frequency analysis approach [13], as follows:

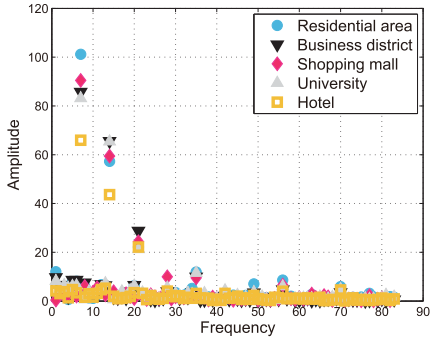


Fig. 5. Frequency analysis of mobile video requests in different locations: amplitude versus frequency.

1) Let $\mathbf{x} = (x_1, x_2, \ldots, x_N)^T$ denote the number of video requests over time, i.e., x_i is the number of requests in time slot i. In our experiments, each time slot is 1 hour, and we study the request samples in 1 week, i.e., $N = 168$.
2) We perform DFT as follows,

$$X[k] = \sum_{n=1}^{N} x_n e^{-2\pi ikn/N},$$

where $X[k]$ is the frequency spectrum of sequence of requests X in the time domain. A larger $X[k]$ indicates that the sequence has a stronger period of k.
3) We study the amplitude of the frequency-domain sequence $X[k]$, in which amplitude and phase represent request volume and their peak-valley time, respectively.

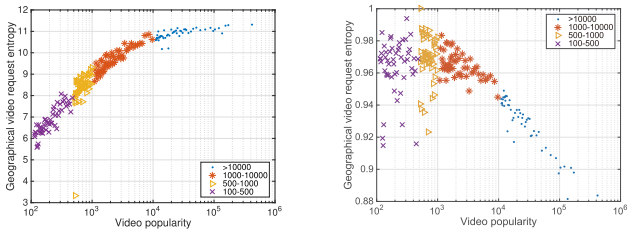
Fig. 5 shows the discrete Fourier transform (DFT) results of the requests over time. In particular, we plot the amplitude versus the frequency of requests in different functionalities of locations. Our observations are as follows: (1) There are some major frequencies with large amplitudes, e.g., $k = 7, 14$, and 21, corresponding to 1 day, 12 hours and 8 hours, respectively. This means that we can use the three frequency components to present the time-domain traffic. Furthermore, we can leverage this property to predict the future traffic. (2) Different functionalities of locations also have different major frequency patterns. For example, the daily pattern is more obvious for the residential areas than the hotels, and the business areas have a strong period of 8 hours. This observation indicates that *the periodical patterns of mobile video requests are highly affected by the functional type of locations*, which can be utilized to distinguish locations with different functionalities.

C. Analysis on Content Video: An Entropy Approach

We study how different videos are actually requested in different locations. To this end, we use an entropy analysis approach. Motivated by the entropy calculation in information theory [9], [14], [38], we define a video request entropy and a location request entropy.

1) Geographical Video Request Entropy: The geographical video request entropy $H^V(v)$ is defined as follows:

$$H^V(v) = -\sum_{l \in \mathcal{L}_v} \frac{n_{vl}}{\sum_{j \in \mathcal{L}_v} n_{vj}} \log \frac{n_{vl}}{\sum_{j \in \mathcal{L}_v} n_{vj}},$$



(a) Non-normalized (b) Normalized

Fig. 6. Geographical video request entropy versus video popularity.

where $H^V(v)$ is the geographical video request entropy for video v, $\mathcal{L}_v$ is the set of locations (e.g., locations defined previously) where video v has been requested, and n_{vj} is the number of requests for video v in location j. A lower value of video request entropy indicates that the video's requests are more diversely distributed across different locations, thereby affecting the caching strategies.

2) Location Request Entropy: The next entropy is location request entropy, which reflects the diversity of videos requested in a particular location. The location request entropy is calculated as follows:

$$H^L(l) = -\sum_{v \in \mathcal{V}_l} \frac{n_{vl}}{\sum_{j \in \mathcal{V}_l} n_{jl}} \log \frac{n_{vl}}{\sum_{j \in \mathcal{V}_l} n_{jl}},$$

where $\mathcal{V}_l$ is the set of videos requested in location l. A larger location request entropy value indicates that the videos requested in the location are more diverse. For the caching strategy, a location with a larger location request entropy generally requires more content items to be replicated to serve the users.

We can compare geographical video request entropy (location request entropy) to evaluate their request patterns given fixed total videos and locations. However, it is unfair to compare geographical video request entropy (location request entropy) directly if the total number of locations where video requests are issued is different (each location has different unique videos). Once any additional location is involved, the entropy will be increased [39]. For example, the requested videos with more locations tend to have larger geographical video request entropy than videos with fewer requesting locations. To overcome this ambiguity, the two entropies have been normalized in our measurement.

3) Entropy Analysis: In this section, we will conduct entropy analysis from two perspectives: video and location.

From the video perspective, we primarily use geographical video request entropy. We first divide videos into four grades according to the requested times (i.e., video popularity) during two weeks, and then we select 50 videos from each grade randomly and compute the geographical video request entropy for these videos. Fig. 6(a) shows that the geographical video request entropy increases as the video popularity increases. This result is consistent with the general understanding. The more popular the video is, the larger its geographical video request entropy is. Consequently, popular videos receive requests from almost everywhere (global distributions), whereas unpopular videos only receive requests from some specific locations (local distributions). Fig. 6(b) shows the

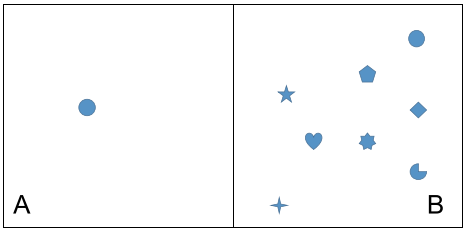


Fig. 7. Location request entropy schemetic diagram.

normalized geographical video request entropy. Interestingly, this figure shows a different result that the more popular the video is, the smaller the normalized entropy is. The reason is that *although the requests for a popular video are requested from more locations, the request distribution of these locations is more skewed compared with that of unpopular videos*.

From the location perspective, we primarily use the normalized location request entropy. We study the distribution of the location request entropies. In Fig. 8(a), we plot the location request entropies $H^L(l)$ of 10514 locations versus the rank of the locations. The results are calculated based on our 2-week traces. We observe that the normalized location request entropy distribution is almost a straight line without the smallest locations, ranging from 0.8 to 1. To better understand these values, Fig. 7 shows the corresponding schematic diagram, where different shapes represent different unique videos. Location A only requests one video each time; thus, the entropy is 0. However, users in location B issue eight requests for eight different videos and the distribution is uniform; thus, the corresponding request entropy is 1. Therefore, the fewer video requests and strongly skewed distributions result in the smallest entropy of the locations. Intuitively, locations with more unique videos have smaller entropy, which have more skewed distributions. Considering cache strategies, *LFU is better for locations with smaller location entropy since there are many different requested videos at each time, whereas LRU is better for locations with larger location entropy*.

We next investigate whether the normalized location request entropy is affected by the characteristics of the location. In particular, we study the correlation between the location request entropy and the number of PoI functionality labels of a location, e.g., residential area. As shown in Fig. 8(b), each sample is the average normalized entropy of locations versus the number of PoI labels of these locations. Our observations are as follows: (1) Locations with a larger number of PoI labels typically have smaller location request entropies, indicating that a location with more "functionalities" has a more skewed request distribution accompanied by more diverse video requests. (2) The relationship approximately follows the quadratic function $y = 0.0003x^2 - 0.0096x + 0.9648$, indicating that the location with a particular number (nearly fifteen) of PoI labels has the smallest entropy.

We also study the impact of the number of users on the location request entropy. In Fig. 8(c), each sample is the average normalized location request entropy versus the number of users requesting videos in these locations. We observe that a larger number of users generally leads to a smaller location request entropy, indicating that *more users generate more skewed request distributions with more diverse content requests*.

Finally, we investigate the impact of user mobility on the normalized location request entropy. We define the user mobility intensity of a location as the mean of all multi-location users who have requested videos in that location. In this experiment, locations with no user movement are not considered. In Fig. 8(d), each sample is the location request entropy versus the user mobility intensity. As shown in this figure, as the user mobility intensity increases from 10^0 to 10^3, the normalized location request entropy gradually increases. We fit the samples into the function $y = 0.0085\log(x) + 0.9273$, implying that user mobility is also a factor for the content diversity. In contrast to user number, the user mobility intensity has a positive impact on location request entropy. *The larger the user mobility intensity is, the larger the location request entropy is and the less the unique video number of the location is.* It is inferred that multi-location users are more likely to request popular videos without increasing the unique video number. One possible reason is that the time of the multi-location users is fragmented such that they are more interested in popular videos. Thus, LFU is more suitable for multi-location users. We will verify the results in Sec. VI.

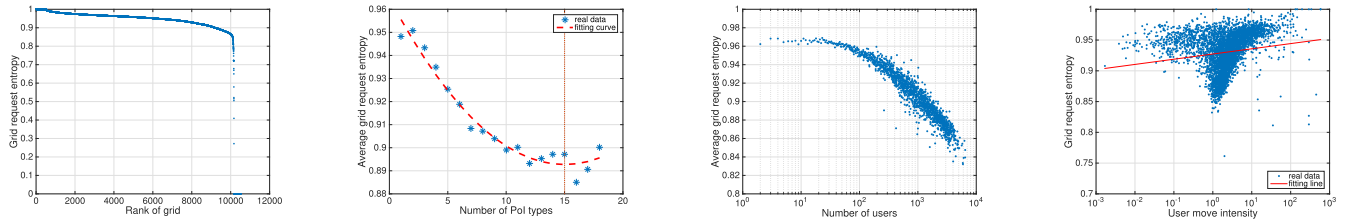
V. Mobile Video Requests Affected by User Mobility Behaviors

In this section, we study what drives the previous request patterns. Particularly, we focus on mobile video user behaviors. In the following experiments on multi-location users, our results are the average results of fourteen days.

A. Mobility Intensity Analysis

In our experiments, we only study the behaviors of *active users* who requested at least ten videos daily in our 2-week traces. Among these 9,576 *active users*, we have 30% *multi-location users* and 70% *simple-location users*, which are defined previously.

1) Movements and Locations Visited: We first study the mobility intensity of the multi-location users. In Fig. 9(a), we plot the fraction of users versus the number of "movements", i.e., the number of requests issued in different locations in *one day*. We observe that the number of movements is generally in the range $[1, 30]$, and the range $[2, 3]$ has the largest fraction of users. The results are quite similar for weekdays and weekends. We next study the number of locations where the requests are issued. In Fig. 9(b), the bars are the fraction of users versus the number of locations where videos are requested in one day. As shown in this figure, as many as 50% of the multi-location users only issued video requests at 2 locations, and 80% of the users only requested videos from less than 4 locations. These results indicate that *it is common for multi-location users to request videos from different locations, but the number of locations (per user) is quite limited*. It provides some basic characteristics to capture the trajectory of multi-location users.



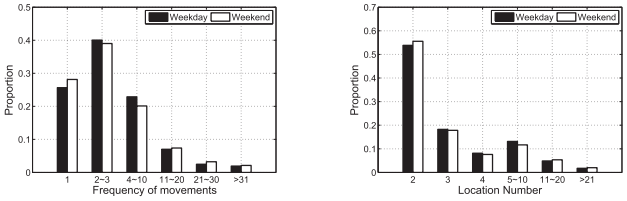
(a) Location request entropy distribution

(b) Location request entropy versus number of PoI labels

(c) Location request entropy versus number of users

(d) Location request entropy versus user mobility intensity

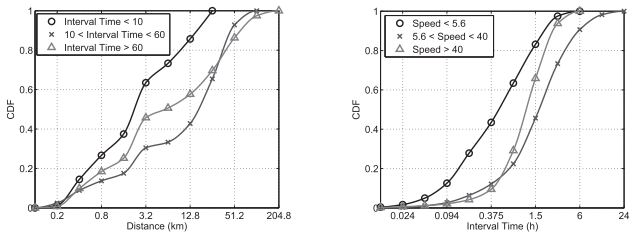
Fig. 8. Content request geographical distribution: entropy analysis.



(a) Movement number

(b) Location number

Fig. 9. Statistics of movement of mobile users.



(a) Distance

(b) Interval

Fig. 10. CDFs of distances and intervals of consecutive requests.

2) Distance and Interval of Movements: We further measure the cumulative distribution of the distances between consecutively visited locations with different time intervals. Fig. 10(a) plots the CDFs of distances between locations where users consecutively request mobile videos. In detail, we select 3 intervals to divide users into the same order: $[0, 10)$ min, $[10, 60)$ min, and $[60, \infty]$ min. We observe that when the interval is shorter than 10 min, the distance is much shorter than that with the other intervals. However, as the interval time increases, the distance does not always become longer. The small time interval indicates that users frequently move between different locations. It is inferred that most users move between 2 or 3 locations in a small time interval.

We also study the intervals between consecutive mobile video requests. In Fig. 10(b), we plot the interval between consecutive requests of users with different movement speeds. We choose two reference speeds: the average speed of walking (i.e., 5.6 km/h) and the average speed of subway (i.e., 40 km/h). As shown in this figure, when the speed is less than 5.6 km/h, most of the request intervals are small. For example, 80% of the request intervals are issued within 1.5 hours, whereas only 40% of the request intervals are issued in 1.5 hours for the movement speed of $[5.6, 40)$ km/h. These observations indicate that the mobility speed of users also affects the request patterns. Moreover these results largely depend on vehicles, which determine the enroute time.

B. Migration Patterns

For the multi-location users who request videos in different locations, we study their migration patterns, i.e., how they move across different locations.

1) Location Migration Pattern: According to our previous observations, users only request mobile videos in a small number of locations. We study how they move across these locations. In Fig. 11, we plot the fraction of users who

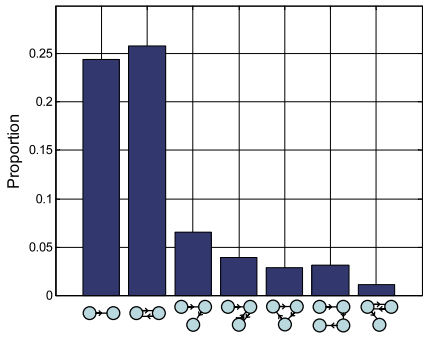


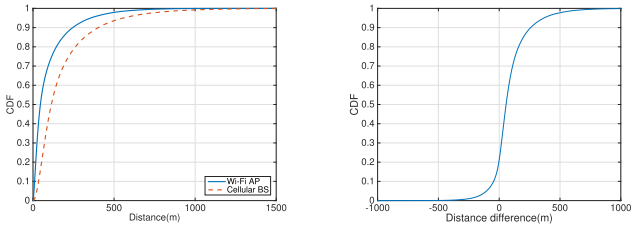
Fig. 11. Fraction of migration patterns.

share the same migration patterns across different locations. In this figure, we plot the most popular 7 migration patterns, which contribute 70% of all the migrations between locations. We observe that *moving between two particular locations constitutes almost* 50% *of the migrations*. Additionally, there are migration patterns across 3 and 4 locations. These results provides us with the basic characteristics to construct connections between different locations for achieving caching cooperation strategies.

2) Location Type Migration Pattern: We study the migration between different functionalities of locations. Based on the PoI information used in our previous measurement studies, we calculate the number of migrations of users from one functionality of location to another functionality of location. As summarized in Table III, each entry is the number of migrations in two weeks, e.g., there are $2,223$ migrations from the hospital areas to the business areas. We observe that (1) it is common for users to move between locations with the same PoI type, e.g., business to business, and (2) there are large migration numbers between some specific pairs of

TABLE III
MIGRATION MATRIX

From / to	Business	Hospital	Resident	Campus	Scenery	Shopping	Hotel
Business	4908	2205	5114	1379	595	1082	657
Hospital	2223	1741	3479	802	394	698	360
Resident	5145	3425	9994	1787	995	1727	907
Campus	1369	797	1743	843	230	367	222
Scenery	596	399	984	215	183	187	123
Shopping	1101	692	1671	358	234	494	169
Hotel	616	367	928	214	114	202	213



(a) CDF of distances between requests and nearest AP/BS

(b) CDF of distance differences between request and Wi-Fi APs and cellular BSes

Fig. 12. Request coverage by edge-network infrastructure.

location functionalities, e.g., the largest migration number occurs between residential areas and business areas.

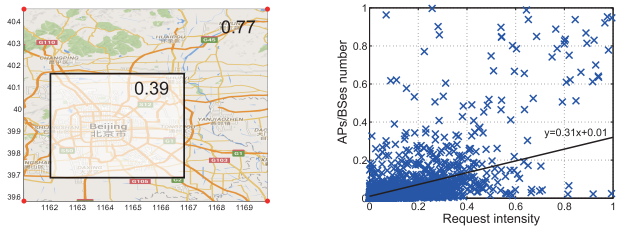
VI. EDGE NETWORK CONTENT DELIVERY FOR MOBILE VIDEO STREAMING

In this section, we compare the effectiveness of Wi-Fi APs and cellular base stations-based edge content delivery solutions, and we discuss the potential improvement to today's wireless networks to enhance mobile video streaming. We first study whether the request intensity in different locations matches the number of edge network infrastructures; we then present the difference between cellular and Wi-Fi on spatial and temporal patterns. In particular, we focus on the effects on caching performance of influencing factors, including different strategies, request density, video and user diversity and user mobility.

A. Request Coverage by Edge Network Infrastructure

To answer the question of whether today's edge network infrastructure can appropriately satisfy the mobile video streaming demand, we measure the distance between users and their nearest infrastructure, and we compare the differences of distribution between requests and edge network infrastructure.

1) Distance between Requests and APs/BSes: We investigate how the mobile video requests can be served by nearby edge network infrastructures, including the Wi-Fi APs (i.e., the smartrouter mode) and cellular base stations (i.e., the femtocell mode). In particular, we study how far away users can find an AP or base station to download videos. Fig. 12(a) plots the CDFs of the distances between the requests and the nearest Wi-Fi APs or cellular BSes that can potentially serve them. We observe that over 95% of the mobile video requests can at least find a Wi-Fi AP within 500 meters or a cellular BS within 750 meters, indicating that edge network video content delivery is promising. We further compare the distance between a video request and the nearest Wi-Fi AP, and the



(a) Selected area

(b) Scatter distribution

Fig. 13. Request intensity versus number of APs/BSes.

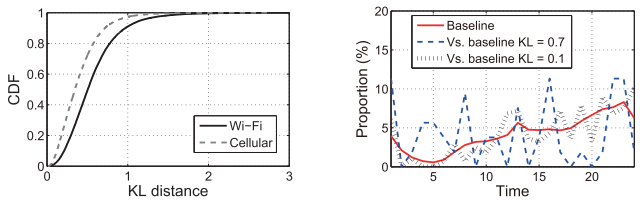
distance between the same request and the nearest cellular BS. In Fig. 12(b), a distance gap larger than 0 suggests that the distance for cellular base station is larger than the distance for Wi-Fi AP. We observe that over 80% of the distance differences are larger than 0, suggesting that Wi-Fi APs are generally closer to users.

2) Request Intensity versus Number of APs/BSes: We also investigate the request intensity (number of requests in different locations) and the number of Wi-Fi APs and cellular base stations in the entire city under two assumptions: all the requests have the same cost, and all APs/BSes have the same power. We use the max-min method to normalize request intensity and number of APs/BSes ranging from 0 and 1. In particular, we investigate whether the requests and the APs/BSes share the same distribution, e.g., there are more APs/BSes if there are more requests in the same location. To this end, we calculate the cosine similarity between the two, i.e., $\mathbf{q} \cdot \mathbf{a}$, where $\mathbf{q}$ is the normalized vector of the numbers of requests in the locations and $\mathbf{a}$ is the normalized vector of the numbers of the APs or BSes at the same locations. A large similarity indicates that the request intensity matches the number of APs/BSes. We observe that the similarity is higher than 0.77, which is considered to indicate a significant similarity. Interestingly, only considering the centralized 30% area that occupies more than 80% of the total requests, the similarity is less than 0.39, as shown in Fig. 13(a). Fig. 13(b) shows the comparison between them. These results imply that a marked difference exists between request intensity and number of APs/BSes, particularly in high-intensity locations where APs/BSes are generally unable to satisfy the requests. It suggests that video service providers should deploy more APs/BSes to better satisfy the users' quality of experience.

3) Wi-Fi/Cellular Stability Analysis: From each Wi-Fi AP and cellular base station perspective, we are interested in the following question. Does the request time distribution of Wi-Fi/cellular follow the global request distribution? To answer this question, we measure the divergence between the time distribution of global requests and single Wi-Fi/cellular requests. To this end, we use the Kullback-Leibler (KL) distance to measure the distance between two distributions, which is defined on two distributions P and Q as follows:

$$D_{KL}(P \parallel Q) = \sum_{t \in T} P(t) \log \frac{P(t)}{Q(t)},$$

where T refers to the set of time, P is the Wi-Fi/cellular distribution of request time on a particular day, and Q is



(a) CDF of KL distance (b) Example

Fig. 14. KL distance between daily request distributions.

the distribution of global request. The KL distance is a non-negative value. It represents the number of extra bits necessary to encode samples from P when using a code based on Q, rather than directly based on P. The smaller the value is, the closer the two distributions are. Fig. 14(a) depicts the CDF of KL distance over Wi-Fi and cellular, which conveys that the daily request distributions of cellular relatively follow the global request distribution. The reason is that Wi-Fi APs attract users mostly from a smaller location where users have particular interests in the content. Fig. 14(b) illustrates two instances of KL distance. It shows that when KL distance equals 0.1, the distribution is similar to the baseline.

B. Performance of Content Caching by Edge Networks

In this section, we evaluate the quality of user experience in current mobile video systems. We assume that a user has a better quality of experience (lower delay) when he is served by the edge cache of a Wi-Fi AP or cellular base station. Thus, we build a discrete trace-driven simulator to evaluate the cache hit rates of conventional caching strategies for Wi-Fi APs and cellular BSes. We simulate mobile video requests, following the records in the real-world traces. We also use the positions of the APs and BSes recorded in our traces to simulate the edge network infrastructure.

1) Experimental Setup: In the simulation experiments, we assume that the average video size S is unit [7], [29], [32] and the evaluation criterion is the total cache hit rate. We use the 2-week records of users' requests to drive the simulation, and we let the requests be served by the nearest Wi-Fi APs or cellular BSes. All the APs/BSes have the same cache capacity C and the default cache capacity is 20 (items). We set the concurrency of APs/BSes as $20/100$ and the bandwidth as $20S/100S$ for APs/BSes to limit the max transmission number per unit time [40]. The radius of each AP/BS is $100m/500m$ [41]. The parameters of the experiments are summarized in Table IV.

In our experiments, we use the following conventional caching strategies: (1) Least recently used (LRU). It discards the least recently used content item first when the cache is full. (2) Least frequently used (LFU). It discards the least frequently used item first when the cache is full. (3) Random replacement (RR). It randomly selects a candidate item and discards it when necessary.

2) Hit Rate versus Capacity: We first study the impact of the cache capacity on the cache hit rate. Fig. 15 shows the cache hit rates of different caching strategies for both Wi-Fi and cellular networks by varying the cache capacity

TABLE IV
SIMULATION PARAMETERS SETTING

Parameter	Value
Average video size S	Unit
Cache capacity C	20 (default)
Concurrency of APs/BSes	$20/100$
Bandwidth of APs/BSes	$20S/100S$
Radius of APs/BSes	$100m/500m$

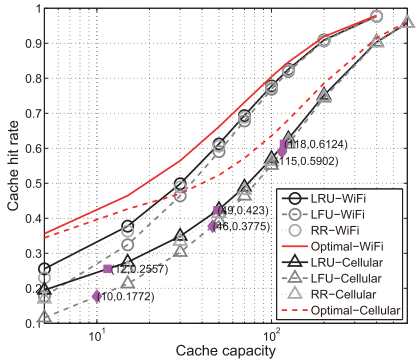
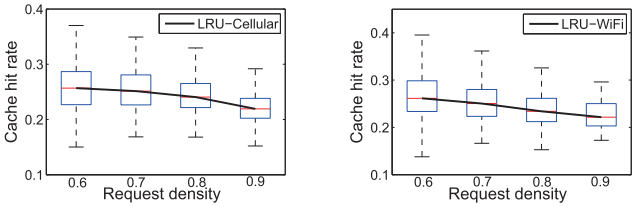


Fig. 15. Cache hit rate under different cache capacity.

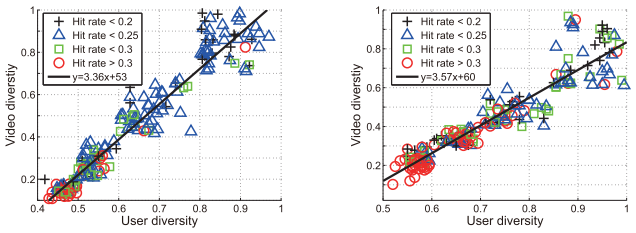


(a) Cellular (b) Wi-Fi

Fig. 16. Cache hit rate versus request density.

from 5 to 600. Our observations are as follows: (1) The cache hit rates in Wi-Fi APs are generally larger than that in cellular BSes, e.g., to reach the same cache hit rate of 0.25 (0.42, 0.61) with LRU, the average cache capacity of the Wi-Fi APs is 5 (20, 50), whereas it is 12 (49, 118) for cellular BSes. Additionally, the result is similar to LFU. (2) LRU, LFU and RR achieve similar cache hit rates, particularly when the cache capacity is large. As the cache capacity increases, the probability of a new item being discarded in RR gradually decreases, resulting in similar performance with LRU. Since the cache contains increasingly more items, all of the strategies achieve high cache hit rates.

3) Impact of Request Density: According to previous measurement studies, different locations present different levels of requests. We study the caching performance for locations with different request density levels. Fig. 16 shows the box-plots of cache hit rates of (a) APs and (b) BSes with different normalized request levels–the request density level is normalized in [0, 1]. We use the LRU strategy with a capacity of 20 unless noted otherwise. We observe only a slight decrease in the cache hit rate with increasing request density, and the variation becomes smaller when there are more requests. These results suggest that the caching strategies are relatively insensitive to the request density.

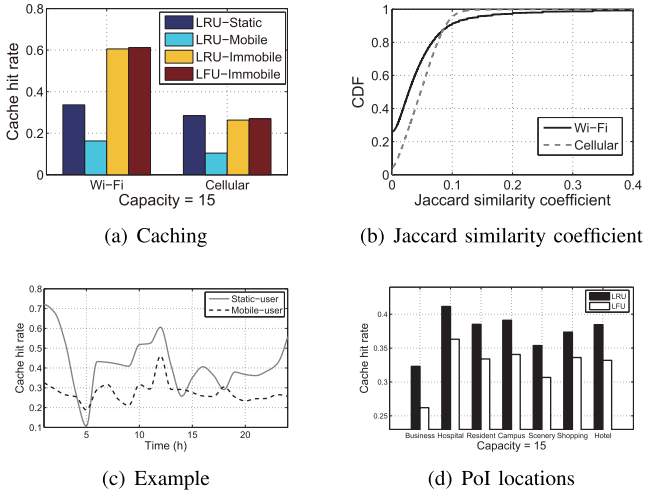


(a) Cellular caching (b) Wi-Fi caching

Fig. 17. Video and user diversity on edge network caching performance.

4) Video and User Diversity: In previous measurement studies, we observe that mobile video request patterns exhibit both content and user diversity. We study the impact of such diversities on edge network content caching performance. We calculate the video diversity as the normalized number of unique videos requested in a location, and the user diversity as the normalized number of users in that location. Fig. 17 shows the cache hit rates with different user diversity and video diversity for Wi-Fi caching and cellular caching, respectively. In Fig. 17(a), we observe that *lower video diversity and user diversity typically lead to higher cache hit rates*, because lower diversities lead to less unique content items requested. We observe different results in Fig. 17(b). For Wi-Fi caching, it is shown that the cache hit rate is considerably higher, and many samples with high cache hit rates appear with large user and video diversities. The result is also consistent with the measurement of location request entropy in Sec. IV-C3. The different fitted lines imply that with the same user diversity, the Wi-Fi APs have larger video diversity on average. For caching strategies, *network designers should deploy larger caches in locations with high video and user diversity to improve the quality of service*.

5) Impact of User Mobility: We study the impact of user mobility on the edge network caching performance. In particular, we investigate the cache hit rates for multi-location users and single-location users. Fig. 18(a) shows that the cache hit rates of multi-location users are always lower than those of single-location users on both LRU and LFU. Thus, the user mobility has a highly negative influence on Wi-Fi/cellular. To determine the possible reasons for why multi-location users have considerably worse caching performance, we first compute the Jaccard similarity coefficient of the users' requested videos in start location l_1 and destination location l_2. The Jaccard similarity coefficient is $J(l_1, l_2) = \frac{|S(l_1) \cap S(l_2)|}{|S(l_1) \cup S(l_2)|}$, where $S(l_1)$ is a set consisting of the videos that users request in location l_1. The coefficient lies between 0 and 1, and the greater the value is, the more similarity they have. Fig. 18(b) depicts the CDF of the obtained similarity coefficients. The majority of location pairs have a similarity coefficient that is less than 0.4, which indicates that the videos requested by users in different locations have greater differences. Second, we assume that the multi-location users are immobile and when moving to l_2 can still fetch content from l_1. The cache hit rates are recorded in Fig. 18(a). Interestingly, the caching performance of Wi-Fi/Cellular is greatly improved, and LFU outperforms LRU, which is an opposite



(a) Caching (b) Jaccard similarity coefficient

(c) Example (d) PoI locations

Fig. 18. Impact of user mobility.

result of single-location users. Thus, *the cache strategies based on LFU are more suitable for multi-location users*. The possible reason is that for destination location l_2, the user becomes a "stranger". Thus, l_2 has difficulty in satisfying the requests from l_1. As with LFU being better than LRU on multi-location users, the result verifies the measurement of location request entropy in Sec. IV-C. Fig. 18(c) presents two instances of single-location and multi-location users. Fig. 18(d) shows that the cache hit rates are different across different functional locations. This indicates that LRU and LFU have different cache hit rates across different locations. Furthermore, the caching performance gap between LRU and LFU is the smallest in shopping areas. One of the reasons is that there are many multi-location users in shopping area. This result also verifies the above observations that LFU is more suitable for multi-location users.

VII. Cache Strategy Based on Measurement Insights

In this section, we design a geo-collaborative caching strategy for mobile video content delivery based on the measurement insights. We also compare its performance with conventional cache strategies.

A. Caching Strategy

Motivated by the measurement insights, we design a geo-collaborative caching strategy for mobile video delivery. Without loss of generality, we consider a general network architecture in which a set $\mathcal{L}$ of L locations provide video content access to their users.

1) Cache Storage: For each location $l \in \mathcal{L}$, we divide the cache storage into 2 parts: one is determined by users residing in the location (single-location users), and the other is determined by multi-location users requesting content there. According to the measurement results, the sizes of the two storage parts are determined by the fraction of the single-location users, i.e., a larger single-location user fraction indicates more storage for content to be requested by users residing in that location.

2) Cross-Location Reference: According to our measurement studies in Sec. IV-B, locations with different functionalities have different request patterns. We propose a geo-collaborative caching strategy as follows. To enable content to be cached by cross-location reference, we propose a rank for locations using the information of user migrations: content requested in a location is referred more if there are more users migrating from/to that location, as follows.

$$r_l^t = M \sum_{i \in \mathcal{L}} o_{il}^{(t-W,t-1)} r_i^{t-1}, \tag{1}$$

where $o_{il}^{(t-W,t-1)}$ is the ratio of the users from location i to l over the total multi-location users in location i in the previous time window $[t-W, t-1]$, W is the time window (one day), and M is a control parameter.

3) Content to Cache: Let $\mathbf{x}_l^t$ denote the strategy to be applied for content replication in location l in time slot t. An entry $x_{lv}^t = 1$ indicates that location l will cache video v, and $x_{lv}^t = 0$ otherwise. Similarly, $\mathbf{y}_l^t$ and $\mathbf{z}_l^t$ represent the caching strategies for single-location users in location l and for multi-location users from other locations, respectively. For $\mathbf{z}_l^t$, we have

$$\mathbf{z}_l^t = \sum_{i \in \mathcal{U}_l} \sum_{j \in \mathcal{L}} r_j^t d_{li}^{(t-W,t-1)} f_{ij}^{(t-W,t-1)} \mathbf{x}_j^{t-1}, \tag{2}$$

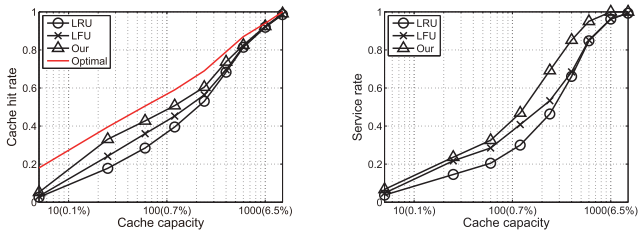
where $d_{li}^{(t-W,t-1)}$ is the fraction of request number of user i over total request number in location l, $f_{ij}^{(t-W,t-1)}$ is the request distribution of user i in location j, and $\mathcal{U}_l$ is the set of users in location l. We iteratively calculate $\mathbf{z}_l^t$ in each time slot.

Caching strategy $\mathbf{y}_l^t$ is determined by the popularity of videos requested by single-location users. For video v, its historical request number before time slot $t-1$ is ρ_v^{t-1}, and it is updated by $\rho_v^{t-1} = \rho_v^{(t-2,t-1)} + e^{-\mu} \rho_v^{t-2}$, where μ is a positive decay factor determined by the video category. To determine $\mathbf{y}_l^t$, location l will cache videos requested with the largest ρ_v^{t-1}. Finally, $\mathbf{x}_l^t$ can be derived by the union of $\mathbf{y}_l^t$ and $\mathbf{z}_l^t$.

B. Performance Evaluation

We use the same simulator from the previous section to evaluate the cache strategy. In the experiments, to ensure the generality that each cellular base station (or Wi-Fi AP) has sufficient requests, only the top 10% most requested cellular BSes (or Wi-Fi APs) are considered.

We first study the impact of cache capacity on the cache hit rate. Fig. 19(a) shows the cache hit rates of different caching strategies by varying the cache capacity from 1 to 1500. As expected, increasing the cache capacity increases the cache hit rate for all the caching strategies, as more requests are satisfied locally without requesting from the CDN servers. Compared with LRU and LFU, the gain of our method increases faster at the beginning. LRU, LFU and our strategy achieve similar cache hit rates when the cache capacity is large. The reason is that when the cache capacity is sufficiently large, each cellular BS can cache all the content and achieve a high cache hit rate.



(a) Cache hit rate (b) Service rate

Fig. 19. Performance comparison (the percentage number in brackets is the ratio of cache capacity to the total number of content).

Next, we study the *service rate*, which is defined as the fraction of the number of users served by APs/BSes over the number of all users. Fig. 19(b) shows the service rates under different cache capacities. Compared with LRU and LFU, the gain of our strategy gradually increases as the cache capacity increases, indicating that the geo-collaborative cache strategy can potentially alleviate the original servers significantly.

VIII. Concluding Remarks

In this paper, we use measurement studies and trace-driven experiments to investigate the performance of edge network content caching for mobile video content delivery. We measure the spatial and temporal request patterns in mobile video systems and the user behaviors that have driven such request patterns. Our results show that the geographic request distribution in a mobile video system can be highly diverse, and the content requested varies among changing locations and periods. Such request patterns are generally determined by user mobility and preference behaviors, in which users exhibit regular commute behaviors, suggesting that joint caching strategies are promising for mobile video content delivery. Next, we compare the effectiveness of cellular and Wi-Fi based edge network caching solutions. Although Wi-Fi and cellular caching are promising, a number of factors including user mobility, content popularity, and cache capacity, have to be taken into consideration for edge network caching for mobile video delivery. Finally, we design a geo-collaborative caching strategy for mobile video delivery based on the measurement insights. Trace-driven experiments further verify the effectiveness of our design.

References

[1] "Cisco visual networking index: Global mobile data traffic forecast update 2014–2019," White Paper, Cisco, San Jose, CA, USA, 2016.

[2] W. Hu and G. Cao, "Quality-aware traffic offloading in wireless networks," in *Proc. 15th ACM Int. Symp. Mobile Ad Hoc Netw. Comput.*, 2014, pp. 277–286.

[3] C. Liang and F. R. Yu, "Wireless network virtualization: A survey, some research issues and challenges," *IEEE Commun. Surveys Tuts.*, vol. 17, no. 1, pp. 358–380, 1st Quart., 2015.

[4] V. K. Adhikari *et al.*, "Unreeling netflix: Understanding and improving multi-CDN movie delivery," in *Proc. IEEE INFOCOM*, Mar. 2012, pp. 1620–1628.

[5] M. K. Mukerjee *et al.*, "Enabling near real-time central control for live video delivery in CDNs," *ACM SIGCOMM Comput. Commun. Rev.*, vol. 44, no. 4, pp. 343–344, 2015.

[6] B. Li, Z. Wang, J. Liu, and W. Zhu, "Two decades of Internet video streaming: A retrospective view," *ACM Trans. Multimedia Comput., Commun., Appl.*, vol. 9, no. 1s, p. 33, 2013.

[7] N. Golrezaei, K. Shanmugam, A. G. Dimakis, A. F. Molisch, and G. Caire, "FemtoCaching: Wireless video content delivery through distributed caching helpers," in *Proc. IEEE INFOCOM*, Mar. 2012, pp. 1107–1115.

[8] M. Ma, Z. Wang, K. Su, and L. Sun, "Understanding content placement strategies in smartrouter-based peer video CDN," in *Proc. ACM SIGMM Workshop Netw. Oper. Syst. Support Digit. Audio Video (NOSSDAV)*, 2016, p. 7.

[9] A. Brodersen, S. Scellato, and M. Wattenhofer, "YouTube around the world: Geographic popularity of videos," in *Proc. 21st Int. Conf. World Wide Web*, 2012, pp. 241–250.

[10] J. Xu, M. V. D. Schaar, J. Liu, and H. Li, "Forecasting popularity of videos using social media," *IEEE J. Sel. Topics Signal Process.*, vol. 9, no. 2, pp. 330–343, Mar. 2015.

[11] A. K. Das, P. H. Pathak, C.-N. Chuah, and P. Mohapatra, "Contextual localization through network traffic analysis," in *Proc. IEEE INFOCOM*, Apr./May 2014, pp. 925–933.

[12] S. Gitzenis, G. S. Paschos, and L. Tassiulas, "Asymptotic laws for joint content replication and delivery in wireless networks," *IEEE Trans. Inf. Theory*, vol. 59, no. 5, pp. 2760–2776, May 2013.

[13] H. Wang, F. Xu, Y. Li, P. Zhang, and D. Jin, "Understanding mobile traffic patterns of large scale cellular towers in urban environment," in *Proc. ACM Conf. Internet Meas. Conf.*, 2015, pp. 225–238.

[14] Z. Li, G. Xie, J. Lin, Y. Jin, M.-A. Kaafar, and K. Salamatian, "On the geographic patterns of a large-scale mobile video-on-demand system," in *Proc. IEEE INFOCOM*, Apr./May 2014, pp. 397–405.

[15] H. Pinto, J. M. Almeida, and M. A. Gonçalves, "Using early view patterns to predict the popularity of YouTube videos," in *Proc. 6th ACM Int. Conf. Web Search Data Mining*, 2013, pp. 365–374.

[16] Z. Li *et al.*, "Watching videos from everywhere: A study of the PPTV mobile VoD system," in *Proc. ACM Conf. Internet Meas. Conf.*, 2012, pp. 185–198.

[17] J. L. Toole, M. Ulm, M. C. González, and D. Bauer, "Inferring land use from mobile phone activity," in *Proc. ACM SIGKDD Int. Workshop Urban Comput.*, 2012, pp. 1–8.

[18] W. Song, D. W. Tjondronegoro, and M. Docherty, "Understanding user experience of mobile video: Framework, measurement, and optimization," in *Mobile Multimedia—User and Technology Perspectives*. Rijeka, Croatia: INTECH Open Access Publisher, 2012.

[19] J. Xue and C. W. Chen, "A study on perception of mobile video with surrounding contextual influences," in *Proc. IEEE 4th Int. Workshop Quality Multimedia Exper. (QoMEX)*, Jul. 2012, pp. 248–253.

[20] D. Ciullo, V. Martina, M. Garetto, and E. Leonardi, "How much can large-scale video-on-demand benefit from users' cooperation?" *IEEE/ACM Trans. Netw.*, vol. 23, no. 6, pp. 1846–1861, Dec. 2015.

[21] D. Ciullo, V. Martina, M. Garetto, E. Leonardi, and G. L. Torrisi, "Asymptotic properties of sequential streaming leveraging users' cooperation," *IEEE Trans. Inf. Theory*, vol. 59, no. 12, pp. 8386–8401, Dec. 2013.

[22] K. Cho, H. Jung, M. Lee, D. Ko, T. Kwon, and Y. Choi, "How can an ISP merge with a CDN?" *IEEE Commun. Mag.*, vol. 49, no. 10, pp. 156–162, Oct. 2011.

[23] Z. Wang, W. Zhu, M. Chen, L. Sun, and S. Yang, "CPCDN: Content delivery powered by context and user intelligence," *IEEE Trans. Multimedia*, vol. 17, no. 1, pp. 92–103, Jan. 2015.

[24] G. Zhang, W. Liu, X. Hei, and W. Cheng, "Unreeling Xunlei Kankan: Understanding hybrid CDN-P2P video-on-demand streaming," *IEEE Trans. Multimedia*, vol. 17, no. 2, pp. 229–242, Feb. 2015.

[25] M. Zhao *et al.*, "Peer-assisted content distribution in Akamai netsession," in *Proc. ACM Conf. Internet Meas. Conf.*, 2013, pp. 31–42.

[26] J. Roberts and N. Sbihi, "Exploring the memory-bandwidth tradeoff in an information-centric network," in *Proc. IEEE 25th Int. Teletraffic Congr. (ITC)*, Sep. 2013, pp. 1–9.

[27] H. Ahlehagh and S. Dey, "Video-aware scheduling and caching in the radio access network," *IEEE/ACM Trans. Netw.*, vol. 22, no. 5, pp. 1444–1462, Oct. 2014.

[28] J. Lin, Z. Li, G. Xie, Y. Sun, K. Salamatian, and W. Wang, "Mobile video popularity distributions and the potential of peer-assisted video delivery," *IEEE Commun. Mag.*, vol. 51, no. 11, pp. 120–126, Nov. 2013.

[29] Y. Zhou, L. Chen, C. Yang, and D. M. Chiu, "Video popularity dynamics and its implication for replication," *IEEE Trans. Multimedia*, vol. 17, no. 8, pp. 1273–1285, Aug. 2015.

[30] M. Garetto, E. Leonardi, and S. Traverso, "Efficient analysis of caching strategies under dynamic content popularity," in *Proc. IEEE Conf. Comput. Commun. (INFOCOM)*, Apr./May 2015, pp. 2263–2271.

[31] M. Leconte, G. Paschos, L. Gkatzikis, M. Draief, S. Vassilaras, and S. Chouvardas, "Placing dynamic content in caches with small population," in *Proc. IEEE Int. Conf. Comput. Commun. (INFOCOM)*, Apr. 2016, pp. 1–9.

[32] J. Hachem, N. Karamchandani, and S. Diggavi, "Content caching and delivery over heterogeneous wireless networks," in *Proc. IEEE Int. Conf. Comput. Commun. (INFOCOM)*, Apr./May 2015, pp. 756–764.

[33] M. A. Maddah-Ali and U. Niesen, "Fundamental limits of caching," *IEEE Trans. Inf. Theory*, vol. 60, no. 5, pp. 2856–2867, May 2014.

[34] K. Poularakis, G. Iosifidis, A. Argyriou, I. Koutsopoulos, and L. Tassiulas, "Caching and operator cooperation policies for layered video content delivery," in *Proc. IEEE Int. Conf. Comput. Commun. (INFOCOM)*, Apr. 2016, pp. 1–9.

[35] J. He and W. Song, "Optimizing video request routing in mobile networks with built-in content caching," *IEEE Trans. Mobile Comput.*, vol. 15, no. 7, pp. 1714–1727, Jul. 2016.

[36] Q. Xu, J. Huang, Z. Wang, F. Qian, A. Gerber, and Z. M. Mao, "Cellular data network infrastructure characterization and implication on mobile content placement," in *Proc. ACM SIGMETRICS Int. Conf. Meas. Modeling Comput. Syst.*, 2011, pp. 317–328.

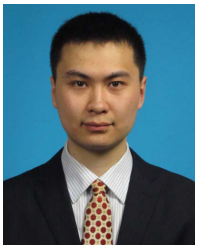
[37] Tencent. *Tencent Wi-Fi*, accessed on Sep. 2016. [Online]. Available: http://www.tencent.com

[38] C. E. Shannon, "A mathematical theory of communication," *Bell Syst. Tech. J.*, vol. 27, no. 3, pp. 379–423, 1948.

[39] L. Chen, Y. Zhou, and D. M. Chiu, "Fake view analytics in online video services," in *Proc. Netw. Oper. Syst. Support Digit. Audio Video Workshop*, 2014, p. 1.

[40] R. Tripathi, S. Vignesh, and V. Tamarapalli, "Optimizing green energy, cost, and availability in distributed data centers," *IEEE Commun. Lett.*, vol. 21, no. 3, pp. 500–503, Mar. 2017.

[41] M. Herlich and S. Yamada, "Optimal distance of multi-hop 802.11 WiFi relays," in *Proc. IEICE Soc. Conf.*, 2014, pp. 44–45.



Ge Ma received the B.E. degree in automation from Tsinghua University, Beijing, China, in 2013, where he is currently pursuing the Ph.D. degree in data science and information technology. His research interests include content delivery network, multimedia big data, caching and prefetching strategy.

Zhi Wang (S'10–M'14) received the B.E. and Ph.D. degrees in computer science from Tsinghua University, Beijing, China, in 2008 and 2014, respectively. He is currently an Assistant Professor with Tsinghua University. His research areas include online social networks, mobile cloud computing, and large-scale multimedia systems. He was a recipient of the China Computer Federation Outstanding Doctoral Dissertation Award in 2014, the ACM Multimedia Best Paper Award in 2012, and the MMM Best Student Paper Award in 2015.

Miao Zhang received the B.E. degree in computer science and technology from Sichuan University, Chengdu, China, in 2015. She is currently pursuing the master's degree in computer science from Tsinghua University, Beijing, China. Her research interests include content delivery network and data analysis.

Jiahui Ye received the B.E. degree in electrical engineering from the Zhejiang University of Technology, Hangzhou, China, in 2016. She is currently pursuing the Ph.D. degree in computer science from Tsinghua University, Beijing, China. Her research interest includes multimedia data analysis and mobile multimedia computing.

Minghua Chen (S'04–M'06–SM'13) received the B.Eng. and M.S. degrees from the Department of Electronic Engineering, Tsinghua University, in 1999 and 2001, respectively, and the Ph.D. degree from the Department of Electrical Engineering and Computer Sciences, University of California at Berkeley, in 2006. He spent one year visiting Microsoft Research Redmond as a Post-Doctoral Researcher. He joined the Department of Information Engineering, The Chinese University of Hong Kong, in 2007, where he is currently an Associate Professor. He is also an Adjunct Associate Professor with the Institute of Interdisciplinary Information Sciences, Tsinghua University. His current research interests include energy systems, such as smart power grids and energy-efficient data centers, intelligient transportation system, distributed optimization, multimedia networking, wireless networking, network coding, and delay-constrained network information flow. He received the Eli Jury award from UC Berkeley in 2007 (presented to a graduate student or recent alumnus for outstanding achievement in the area of systems, communications, control, or signal processing) and The Chinese University of Hong Kong Young Researcher Award in 2013. He also received several best paper awards, including the IEEE ICME Best Paper Award in 2009, the IEEE TRANSACTIONS ON MULTIMEDIA Prize Paper Award in 2009, and the ACM Multimedia Best Paper Award in 2012. He is currently an Associate Editor of the IEEE/ACM TRANSACTIONS ON NETWORKING. He serves as the TPC Co-Chair of ACM e-Energy 2016 and the General Chair of ACM e-Energy 2017.

Wenwu Zhu (S'91–M'96–SM'01–F'10) received the Ph.D. degree in electrical and computer engineering from the New York University Polytechnic School of Engineering, New York, NY, USA, in 1996. He was a Senior Researcher and a Research Manager with Microsoft Research Asia, Beijing, China. He was the Chief Scientist and the Director with Intel Research China, Beijing, from 2004 to 2008. He was with Bell Labs, Murray Hill, NJ, USA, as a member of Technical Staff from 1996 to 1999. He is currently a Professor and the Deputy Head of the Department of Computer Science, Tsinghua University, Beijing. His current research interests include the areas of multimedia computing, communications, and networking. He is a AAAS Fellow, an SPIE Fellow, and an ACM Distinguished Scientist. He was a recipient of Best Paper Awards, including T-CSVT in 2001 and ACM Multimedia 2012. He has been serving as the Editor-in-Chief of the IEEE TRANSACTIONS ON MULTIMEDIA (TMM) since 2017. He served on the Steering Committee of TMM in 2016 and the IEEE TRANSACTIONS ON MOBILE COMPUTING (TMC) from 2007 to 2010. He has served on various Editorial Boards, such as a Guest Editor for the Proceedings of the IEEE, the IEEE JOURNAL OF SELECTED AREAS IN COMMUNICATIONS, the IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS FOR VIDEO TECHNOLOGY (T-CSVT), and an Associate Editor of TMM, the *ACM Transactions on Multimedia, Communications, and Applications*, T-CSVT, TMC, and the IEEE TRANSACTIONS ON BIG DATA.