Spatial Popularity and Similarity of Watching Videos in a Large City

Huan Yan†, Jiaqiang Liu†, Yong Li†, Depeng Jin†, Sheng Chen‡
† Department of Electronic Engineering, Tsinghua University
‡ School of ECS, University of Southampton
Email: liyong07@tsinghua.edu.cn

***Abstract*—With the popularity of watching mobile videos, many works focus on the geographic features of user viewing behaviors, but few study them in the context of an entire metropolitan city. Different regions of a large city have different intensity of economy activities with respect to their different distances to the downtown, and how this will influence video popularity and similarity is still unclear. To quantitatively study the spatial popularity and similarity of watching videos in a large urban environment, we collect a dataset with two-month video view requests from the largest network provider in Shanghai, containing top six content providers, and study the spatial features of video access in regions of different scales. We find that 1) video popularity and similarity exist at different scales of city division; 2) the concentration of video popularity becomes higher as the region is closer to downtown; 3) when comparing the regions of same scale, the similarity of popular videos becomes lower as the region is farther away from the downtown. Finally, we correlate our findings with cache deployment, advertising and video recommendation to illustrate the implications.**

I. Introduction

In recent years, the advance in mobile network technologies and the proliferation of smart mobile devices enable a large number of users to conveniently watch mobile videos at any time and any place, leading to a sharp increase of mobile video traffic. According to Cisco's survey [1], the mobile video traffic will account for 72% of the global mobile traffic by 2019, which is 13 times larger than that of 2014. To cope with this explosive growth of mobile video traffic, network providers have to take effective actions to optimize network performance and improve capacity. Under such context, understanding the spatial popularity and similarity of watching videos with the evolution of mobile networks is particularly important. Despite that recent works [2], [3] have studied the geographic popularity of video views in the world and country scale, characterizing them in a large city environment remains unexplored. Specifically, it is unclear that how the popularity and similarity of watching videos are at different scales of a city. How the popularity and similarity of watching videos are in different regions of a city is also unknown. These issues are important for the network providers to optimize mobile networking such as cache deployment. Obtaining the relationship between different regions and user behaviors helps explore whether the activities of economy will influence the popularity and similarity of user views, which also benefits content providers, enabling them to provide better video service for example.

Thus, associating with different scales and different regions in a metropolis area, we focus on two key problems that need to be investigated:

Problem I: How is the video popularity distributed? At different scales, due to diverse user interests, the concentration of video popularity could be different. Thus, the influence of the region size needs to be investigated. On the other hand, between the downtown and the suburb, whether there exist the differences of the concentration is also a key consideration.

Problem II: How is video request similarity distributed? Under a given scale, how to characterize video request difference between different regions? Meanwhile, what is the difference of the similarity of user views between the downtown and suburb? These are the issues that must be addressed.

In this paper, we aim to systematically study the above problems by analyzing the two-month real-world video request dataset, collected at the gateway of a major Internet service provider (ISP) in one of the largest cities in China, Shanghai. The data contain more than 200 million view requests for videos from top six most popular content providers. We divide the city into non-overlapping regions of four scales by using cellular network infrastructure of the same ISP. At the same time, we exploit topic, defined as videos of episodes/clips in the same program, as a basic unit of video measurement by aggregating videos from different content providers. Then, we discuss spatial features of watching videos related to the above introduced two problems by defining metrics of popular topic number, view concentration and popular topic similarity to characterize them. We carry out a thorough analysis based on these metrics, and our major findings corresponding to the two problems can be summarized as follows:

- We validate the well-known Pareto principle of video popularity at the city scale. Furthermore, we show that the Pareto principle exists in any size of scales studied, which indicates that video popularity exists at any scale. Also, we observe that the concentration of video popularity becomes higher as the size of scale increases, and it becomes higher near the downtown than in the suburb. In particular, associating with different video types such as movies, cartoons, show and TV series, we find that in the downtown the concentration on TV series and shows is higher while opposite is true with the other two types.
- We observe that the similarity between the sets of popular videos of any two regions of small scale is low, but the similarity increases with the scale. Also, the regions nearer the downtown have much more similar popular video requests than the region in the downtown. Furthermore, between the regions in the downtown and in the suburb, the similarity of movies is lowest while the similarity of shows is highest.

Finally, we exploit our findings for potential applications on cache deployment strategies, advertisement and video recommendation. Taking the cache deployment as an example, according to our results, we need cache different contents and number of topics with specified types, e.g., the topics belonging to TV series should be cached more in the downtown than in the suburb.

The rest of this paper is organized as follows. Next, we describe the datasets and basic data processing. We then introduce the methodology and metrics. We analyze the concentration and similarity of video popularity, and then discuss

TABLE I
THE STATISTICS OF VIEW REQUESTS DATA.

	TV series	Show	Movie	Cartoon
# of topics	7942	2238	10797	2570
# of videos	173749	51952	35459	58186
# of views	141559110	26409539	26613772	13805269

the obtained results. After that, we discuss the implications of our findings. After the survey of related works, we finally conclude the paper.

II. DATA COLLECTION AND PROCESSING

A. Data Collection

In this paper, we analyze the logs of deep packet inspection appliances deployed at the service gateways of a major ISP network in Shanghai, one of the largest cities in China. The logs consists of individual entries, where each entry represents a view requests for a specific video, and records the information of user ID, location, and URL for the request. We further obtain the context of the requested video, including its content provider, ID, type, and description of name, by crawling the URL from web. From the perspective of ISP, this dataset is comprehensive to study the spatial popularity and similarity of watching videos in a city, because all view requests, despite of the content providers, need to go through the ISP gateway. Overall, our dataset is collected during Nov. 1 and Dec. 31, 2014, and it covers 200 million view requests, 1.4 million users, and 1 million videos from top six most popular content providers in China.

In addition, as we aim to investigate the influence of the region size on the concentration and similarity of video popularity, we further collect information of mobile network infrastructure in Shanghai by the same ISP, including the ID, location and associated district of each base station (BS), base station control (BSC) and mobile switch center (MSC). Combining these two datasets of logs of view requests and mobile network structure enables us to characterize the user interests on specific video topics in a specific region and study the spatial features of video popularity and similarity under different scales.

B. Aggregating Videos by Topic

Videos in our dataset come from six content providers, and some of them could be duplicated or aliased, which should be aggregated to eliminate their inference on popularity analysis. Besides, instead of a single video, e.g., one episode of a TV series, we are more interested in the spatial popularity of a set of videos with the same topic, e.g., videos of the same TV series. To this end, we aggregate videos by topic, defined as a set of videos or episodes and/or clips in the same TV series, movie, show or cartoon program. Specifically, we crawl the program lists of TV series, cartoon, show and movie from the websites of six content providers. We then use the title of each program as the keyword of one topic, and classify the videos into topics by matching these keywords with the name strings.

Finally, we obtain 23463 topics, which covers 84% of overall videos [1]. Table I summarizes some basic statistics of the topics by the type of TV series, show, movie, and cartoon. To illustrate that the topic provides an appropriate granularity to aggregate videos, we plot the cumulative distribution function (CDF) of the number of videos in each topic in Fig. 1. We can

[1]"The rest of videos cannot be classified because their names are usually about the content of the episodes and/or clips and do not contain the title of the corresponding program.These videos are removed in our following analysis."

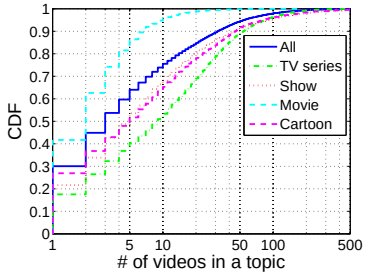
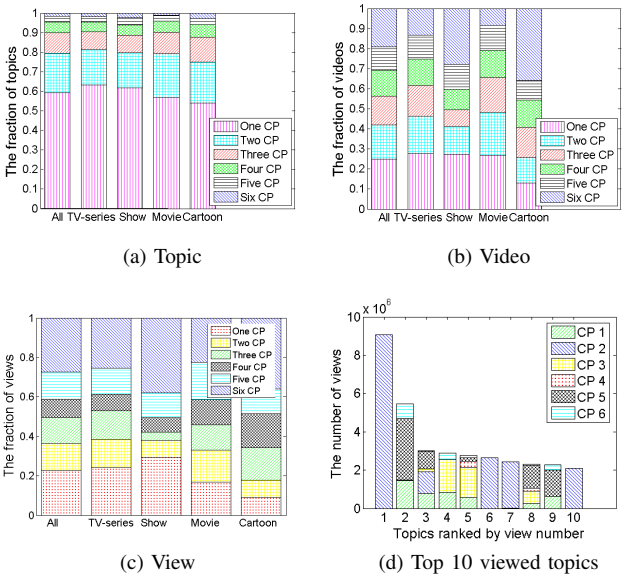


Fig. 1. The CDF of the number of videos per topic.

observe that 90% of topics have less than 25 videos, which suggests that the topic provides a moderate granularity for the aggregation. Overall, the topic of movie usually has the least number of videos, while the topic of TV series usually has more videos due to a large number of episodes. This is in accordance with our intuition and implies the correctness of the aggregation.



(a) Topic (b) Video

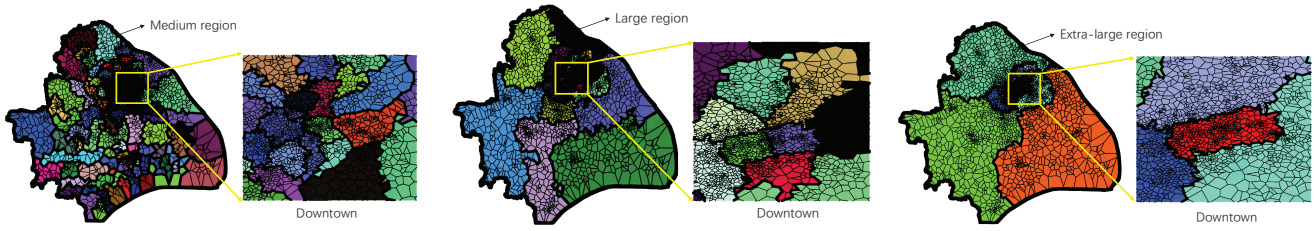
(c) View (d) Top 10 viewed topics

Fig. 2. The fraction of topics, videos, and views occupied by six classes of topics that involve different numbers of content providers (CP) ((a)∼(c)) and the number of views from different content providers for top 10 viewed topics (d).

We further study the involvement of content providers in each topic to illustrate the necessity to aggregate the videos from different providers. Specifically, we first classify the topics into six classes according to the number of providers they involve, e.g., class Two CPs mean that the videos in this topic come from two content providers, and then calculate the fraction of topics, videos, and views that each class accounts for. As shown in Fig. 2, we observe that about 40% topics cover at least two providers, and these topics account for more than 70% videos and about 80% views, which suggest that from the perspective of ISP, the videos offered by any single provider are insufficient to capture user behaviors on a specific topic. This observation can be further confirmed by analyzing the content providers involved in the requests for the top viewed topics: as shown in Fig. 2 (d), six of top ten most viewed topics involve at least three providers, and the views of top 5 topics cover all six providers. Overall, these statistics indicate that it is essential to aggregate videos from multiple providers and the topic is an appropriate granularity to study video popularity and similarity.

C. City Division

To study the spatial popularity and similarity of watching videos at different scales, we divide the city into multiple non-



(a) Medium scale (b) Large scale (c) Extra-large scale

Fig. 3. The geographic distribution under different scales.

overlapping regions with different sizes. Specifically, as the view requests are first sent to the BSs or access points (APs), we obtain the coverage area of each BS/AP and regard them as basic regions. Then we build larger regions by aggregating adjacent basic regions according to the structure of the ISP's network infrastructure.

As for each base station we only know its location, we use Voronoi diagram to obtain their coverage areas. Formally, given the locations of N BSs', denoted by $\{l_1, l_2, \cdots, l_N\}$, the Voronoi diagram gets the coverage areas for the BSs as $\{R(l_1), R(l_2), \cdots, R(l_N)\}$, where $R(1_i)$ denotes the coverage area of the i-th BS, and any location p inside $R(l_i)$ satisfies that the Euclidean distance between p and l_i is smaller than that between p and l_j for any $l_j \neq l_i$.

TABLE II
THE COMMUNICATION ENTITIES OF THE MOBILE NETWORK CORRESPONDING TO THE DEFINED SCALES.

Small scale	Large scale	Extra-large scale
BS/NodeB/ Evolved NodeB (eNodB)	BSC/ Radio Network Controller (RNC)	MSC/ Serving GPRS Supporting Node (SGSN)/ Serving Gateway (S-GW)

To obtain larger regions, we aggregate coverage areas of adjacent base stations by referring to the cellular network infrastructure. As summarized in Table II, the entity in a larger scale controls multiple entities in the smaller scale. For example, a MSC controls multiple BSCs, while a BSC controls multiple BSs. Based on these relationships, we aggregate BSs belonging to the same entity to form larger regions. Besides, to make our study more comprehensive, we also construct larger regions by aggregating BSs belonging to the same geographical district. Overall, we divide the city by four scales: the division using BS's coverage area as the small scale, and the division by aggregating adjacent BSs in the same district as medium scale, the division by aggregating BSs controlled by the same BSC and MSC as the large and extra-large scales respectively. Finally, there are 93, 18 and 7 regions in the medium, large and extra-large scales, respectively. Fig. 3 shows their geographic distribution, where a smallest region represents the coverage area of a BS, and the adjacent smallest regions with the same color represents that they are aggregated into the same larger region.

We can divide the city by four scales of different sizes using above approach. To validate the rationality of such division, we focus on two questions: whether the number of the aggregated views in regions of one scale is significantly different to another scale; and whether view requests are concentrated in a small fraction of regions under each scale. To answer above questions, we plot the overview view distribution under four scales in Fig. 4. We observe that the number of views in a region of different scales exhibits different orders of magnitude, i.e., from hundreds to hundreds of thousands on average. Besides, for regions of the same scale, the number of views mostly concentrates on specific range, as the difference between 25% and 75% percentile of views where it distributes is small, i.e., the number of views in the small scale is ranging from 162 to 992.

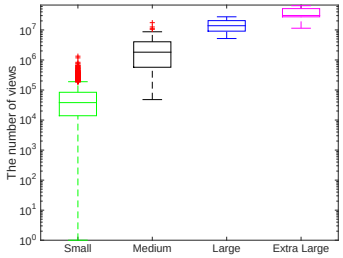


Fig. 4. The number of views in regions of different scales.

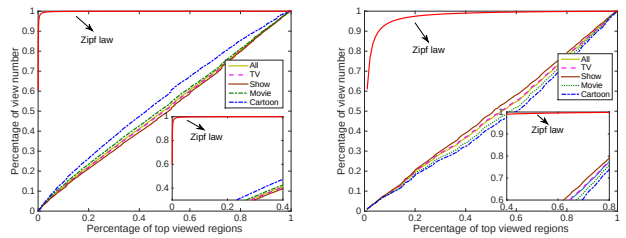
To quantitatively analyze view distribution over regions of the same scale, we calculate the fraction of views accounted by different fractions of regions that attracts highest views. We show the results in Fig. 5. From the results, we observe that the top 20% of regions only generate 20% of video views in the small scale, which is 5 times smaller than the Zipf's law that denotes the standard geographical concentration [13] We can observe similar gaps in other scales from Fig. 5 (b)-(d), which implies that the views do not exhibit geographical concentration under different scales. This observation also stands when we count the number of views for different video types separately. These results demonstrate that the views for different video types also do not exhibit the geographic concentration.

Combining the above two aspects, we conclude that dividing the city at the defined four scales provides a good method to study the spatial popularity and similarity of watching videos.

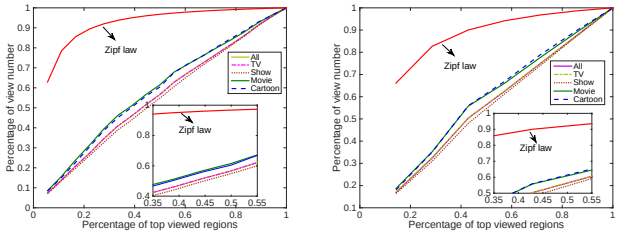
III. METHODOLOGY AND METRICS

In order to quantitatively answer the two problems regarding the spatial characterizations of video popularity and similarity, we investigate the view patterns related to them. Before introducing the methodology and related metrics, we first describe the notations utilized below.

We denote the topic set as $\{p_1, p_2, \cdots, p_M\}$, where M is the number of topics. We denote n_i as the number of topics that receive at least one view in region i. For each region, we rank the topics in descending order by the number of views, and denote $u_{i,k}$ as the number of users viewing the kth rank topic, i.e., the kth most viewed topic, of region i, and $v_{i,k}$ as its view number. We denote $v_i = \sum_{j=1}^{M} v_{i,j}$ as the total number of views of region i. We are now ready to discuss the methodology utilized in our study.



(a) Small scale (b) Medium scale



(c) Large scale (d) Extra-large scale

Fig. 5. The percentage of view number by each user as a function of the percentage of top viewed regions.

A. Problem I: Video Popularity Concentration

Since users in some regions may be interested in only a few popular topics, while for other regions video popularity may cover a large number of topics. Therefore, in order to study the concentration of video popularity, we investigate the distribution of the number of popular topics over regions.

We exploit the view number as a measure of popularity, and define the popular topics as the ones that receive more than a given number of views. Formally, given a threshold of view number, say n, we define ***popular topic number*** of region i, denoted by $P_i(n)$, as follows:

$$P_i(n) = \sum_{j=1}^{M} I(v_{i,j} - n), \tag{1}$$

where $I(x) = 1$ if $x \geq 0$ and otherwise $I(x) = 0$.

On the other hand, to eliminate the influence of the specified view number, we use the fraction of the top viewed topic to define popular topic. Specifically, given f $(0 < f < 1)$, we regard the popular topics of region i as the most viewed topics that account for at least f of the total topics. Formally, we define the popular topic number $Q_i(n)$ of region i given f as follows:

$$Q_i(f) = \lfloor f \times n_i \rfloor. \tag{2}$$

A higher value of $P_i(n)$ or $Q_i(f)$ indicates that region i has more popular topics.

In general, if a large fraction of views concentrates on a few popular topics, it represents the high concentration of video popularity. On the other hand, if the views are uniformly distributed to many different topics, the concentration is low. Therefore, in order to study the spatial characterization of video popularity, we investigate the concentration of views on popular topics under each scale. To quantify it, we define ***view concentration*** of popular topics of region i as follows:

$$\begin{aligned} C_i^P(n) &= \frac{1}{v_i} \sum_{j=1}^{P_i(n)} v_{i,j}, \\ C_i^Q(f) &= \frac{1}{v_i} \sum_{j=1}^{Q_i(f)} v_{i,j}. \end{aligned} \tag{3}$$

In the above definition, a higher value of $C_i^P(n)$ or $C_i^Q(f)$ indicates a higher view concentration and, consequently, a higher concentration of video popularity.

To eliminate the influence of user number on the number of views in a region, we define average views of popular topics by each user in region i as follows:

$$A_i(f) = \frac{1}{Q_i(f)} \sum_{j=1}^{Q_i(f)} v_{i,j}/u_{i,j}. \tag{4}$$

According to the above definition, a higher value of $A_i(f)$ indicates the higher view concentration of video popularity by each individual user.

B. Problem II: Video Request Similarity

Video requests in a region concentrate on the most popular topics. If different regions have similar set of popular topics, there is no significant difference with video popularity. Therefore, in order to answer the question of how similarity of popular video requests is, we study the similarity between the popular topics of different regions. Note that Jaccard similarity coefficient is a well defined measure to characterize the similarity between two sets, specifically, it equals to 1 if the two sets are equal, and it is smaller than 0.5 if they have low similarity. We also use Jaccard similarity coefficient to quantify the similarity between different sets of popular topics. Let Φ_i^k denote the set of top k ranked popular topics of region i. Then, we define ***popular topic similarity*** of region i and region j, where $i \neq j$, as follows:

$$J_{i,j}^k = \frac{\mid \Phi_i^k \cap \Phi_j^k \mid}{\mid \Phi_i^k \cup \Phi_j^k \mid}. \tag{5}$$

According to the definition, users in regions i and j have different popular topic requests if $J_{i,j}^k$ is small.

IV. Spatial Popularity and Similarity Analysis

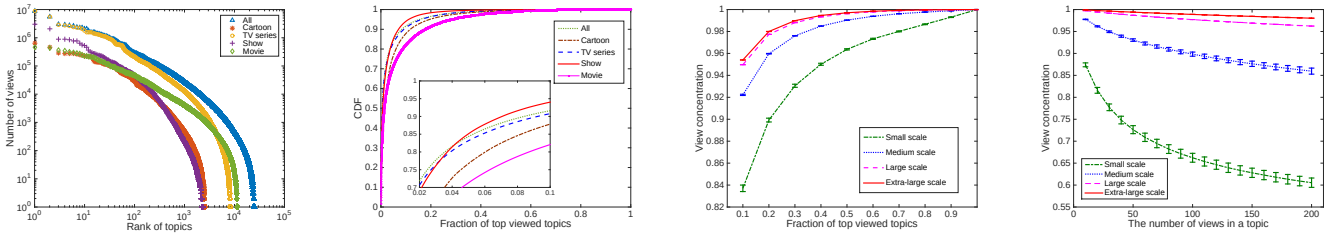
In this section, we study the view patterns by utilizing the introduced methodologies and defined metrics with the purpose to answer the two key problems about spatial popularity and similarity of watching videos.

A. Concentration of Video Popularity

In order to study Problem I of how concentration of video popularity is in regions of four scales. We investigate the popularity distribution of topics and user view concentration on video topics.

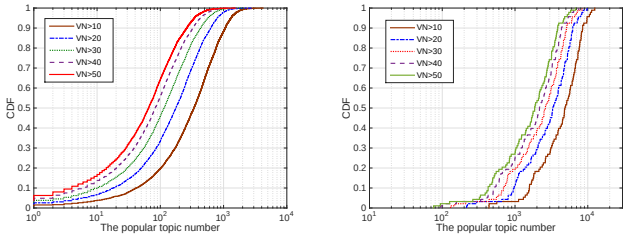
We first consider the overall topic popularity distribution in the city. As depicted in Fig. 6 (a), the number of views on a topic exhibits a strong log-log linear relationship with its rank, indicating the power law distribution of topic's popularity, i.e., the number of views on the topic with rank k is proportional to $k^{-\alpha}$. The power law distribution of popularity reveals that most views concentrate on the most popular video topics. To further verify this, we calculate the percentage of views as a fraction of top viewed topics and plot the result in Fig. 6 (b), where we can observe that top 10% topics accrue more than 80% of views for all different types of TV series, show, movie and cartoon. When we consider all types together, the fraction of views on top 10% topics is larger than 90%. These results validate the well-known Pareto principle of video popularity that describes the concentration of video popularity towards a few popular topics at the city scale.

To study the concentration of video popularity under four different scales, we study the view concentration under these scales according to (3). As shown in Fig. 6 (c), in the small and medium scales, top 10% of popular topics attract more then



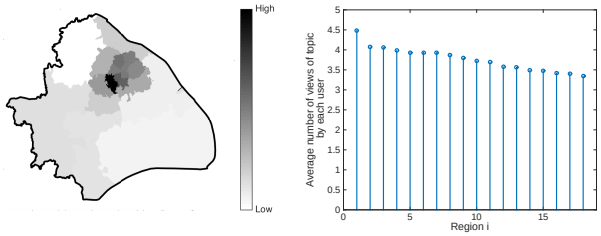
(a) Number of views per topic. (b) CDF of the percentage of views in the city. (c) Average view concentration $C_i^Q(f)$ under different scales. (d) Average view concentration $C_i^P(n)$ under different scales.

Fig. 6. The distribution of the number of views per topic and the view concentration under different scales.



(a) Small scale (b) Medium scale

Fig. 7. The popular topic number P_i as a function of the number of views (VN $>$ 40 means that n in (1) equals to 40).



(a) Large scale (b) Average views

Fig. 8. The spatial distribution of views of top 30% topics by each user in large regions.

80% of views, while in the large and extra-large scales, top 10% of popular topics have more than 90% of views. We find that the results conform to the Pareto principle under different scales. Similarly, in Fig. 6 (d), we observe that in the medium scale, popular topics whose view number exceeds 200 attract more than 80% of views, while in the small scale, popular topics whose view number exceeds 50 have more than 80% of views. Therefore, we validate that *the Pareto principle of video popularity applies to different scales*. Thus, *the larger scale implicates the higher concentration of video popularity*.

We then study the distribution of the popular topics of regions of the same scale to explore their spatial distribution. We measure the popular topic number according to the definition (1) and show the results under small and medium scales in Fig. 7. From the results, we can observe that the number of popular topics exhibits a wide range distribution. Specifically, there is a gap of more than 10 times between the smallest and largest popular topic numbers for regions in both small and medium scales. This wide range distribution shows that each region has different numbers of popular topics and there exists significant differences in the concentration of video spatial popularity.

To understand the concentration of video popularity in the spatial dimension of the city, we calculate the average number of views of top 30% topics by each user according to (4) in the large scale and plot the results in Fig. 8. As shown in Fig. 8 (a) where the darker color represents the larger number of views, we find that the average views increase from the suburb to the downtown, and reach the peak in the downtown. This shows

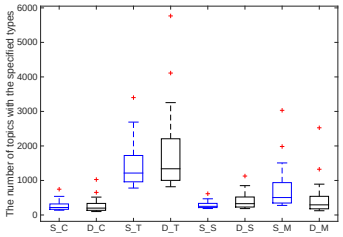


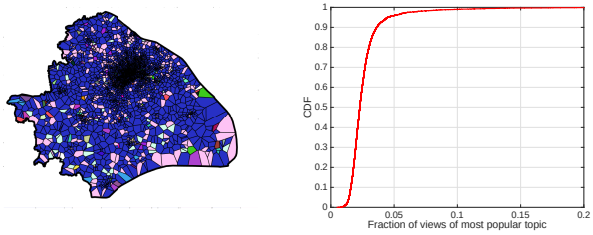
Fig. 9. The number of popular topics whose number of views is range from 50 to 1000 with different types between suburb and downtown.

that video popularity exhibits higher concentration near the downtown. Fig. 8 (b) plots the average views per topic by each user in each region, which reflects the concentration of video popularity for an individual user. To investigate the difference of video popularity between the suburb and downtown, we select two regions corresponding to these two locations where the numbers of users are approximately the same and calculate the number of popular topics whose number of views is range from 50 to 1000 with four types. In Fig. 9, the terms S_C, S_T, S_S and S_M denote the topics with cartoon, TV series, show and movie in the suburb, respectively, while D_C, D_T, D_S and D_M represent the topics with cartoon, TV series, show and movie in the downtown, respectively. From the results, we observe that the number of topics with TV series and shows in the downtown is larger than that in the suburb, while the topics with cartoon and movie attract higher concentrations in the suburb than that in the downtown. One of the reasons is that the more diversity of TV series and shows can attract more preferences across all ages, especially in the downtown. As a result, there are great differences of video popularity in terms of the topic types between the suburb and downtown in the city, and hence implicates that different regions corresponding to different intensity of economy activities (e.g., most economically developed regions locate in the downtown) have a great impact on video popularity.

B. Similarity of Video Requests

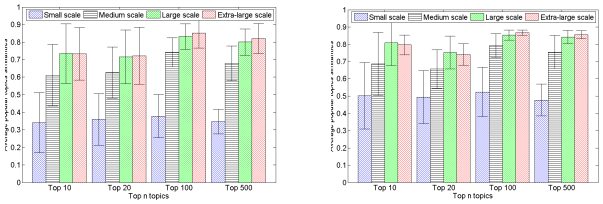
Now, we focus on Problem II of how similarity of video requests is. To address it, we leverage the metric of popular topic similarity to analyze whether the popular topics are the same in different regions.

We select the topics with the most views in the regions of small scale and plot their geographic distributions in Fig. 10 (a), where different color represents different most popular topics. As can be seen from Fig. 10 (a), there are no significant differences in colors of geographically adjacent regions, which demonstrates that the most popular topics are the same over most of regions and thus users in these regions take the same interest in some topic. However, the percentage of views of the most popular topics in these regions is relatively low, as shown in Fig. 10 (b). This shows that even



(a) Small scale (b) CDF

Fig. 10. Illustrating the diversity of most viewed topics in small regions.



(a) Four-scale local comparison (b) Four-scale global comparison

Fig. 11. The popular topic similarity (a) compared with that of locally popular topics under four scales; (b) compared with globally popular topics under four scales.

though some popular topic attracts the most views in most of regions, we cannot neglect the influence of other popular topics that reflect the diversity of video popularity in these regions.

To quantitatively study the diversity of popular topics between regions in different scales, we choose different number of popular topics in each region to calculate their similarities according to (5). Fig. 11 (a) depicts the average similarity with different numbers of popular topics, where we observe that higher similarity exists in larger scales. For example, the similarity of top 10 topics in the extra-large scale is 0.75, which is about 2 times of that in the small scale. This is because the popular topics exhibit the characteristic of concentration in larger regions. To further illustrate this characteristic, we select the most viewed topics in the city and compute their similarity with most popular topics in each region. We show the results of similarity in Fig. 11 (b). From Fig. 11 (b), we observe different similarities under different scales, where the larger the scale is, the higher the similarity can be observed, e.g., the average similarity of top 10 topics under the larger scale is close to 0.8, while it reaches 0.5 under the small scale.

The above analysis focuses on the similarity of popular topics in regions of different scales. To investigate the similarity of popular topics between the downtown and suburb, we choose the concentration of the downtown, the region filled with red color in Fig. 12, as the baseline and calculate the similarity of top 15 topics between this region and other regions in the large scale. As shown in Fig. 12, where the darker color represents higher similarity, we find that the similarity of popular videos becomes low from the downtown to the suburb. Specifically, we choose a region in the suburb and calculate the similarity of popular topics that have more than a given number of views in terms of video types. As shown in Fig. 13, the similarity of topics belonging to movies is lower than that belonging to cartoons, shows and TV series, e.g., the similarity of the popular topics that belong to movies and have more than 200 views is only about 0.2, which is the lowest compared with the similarity of topics belonging to other three types. Further, the topics belonging to shows account for the highest similarity when the number of views in a topic is smaller than 800, which reveals users in these two regions favour similar topics belonging to shows.

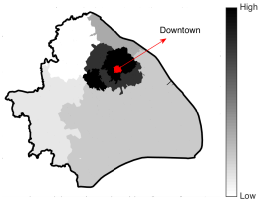


Fig. 12. The spatial distribution of similarity of top 15 topics in large regions compared with that in the downtown.

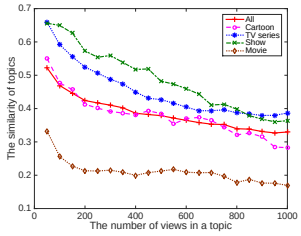


Fig. 13. The similarity of popular topics with different types between suburb and downtown.

In summary, we observe that *great difference of popular topics exist at different scales*. Especially, *the similarity of popular videos is low in the small scale, while it increases with the increase of the scale*. Furthermore, *the nearer the region is to the downtown, the higher the similarity is*. At the same time, *users request the similar popular topics belonging to the show and different topics related to the movie between the downtown and suburb*.

C. Summary of Our Findings

By analyzing the spatial characterization of video popularity and similarity in the city, we have obtained some interesting and important findings, which are summarized as follows:

1) Under the studied four scales, most views are concentrated on a few popular topics, and the concentration of video popularity becomes higher in larger scale. Under a given scale, the concentration of video popularity exists great difference in terms of the number of popular topics in the regions.

2) Considering region locations' influence on the concentration of video popularity, we find that higher concentration exists in the downtown. Video popularity on TV series and shows have higher concentration in the downtown, while the higher concentration of video popularity on cartoons and movies exists in the suburb.

3) The popular topic similarity between different small-scale regions is low, and it increases with the scales. Considering region locations' influence on the similarity of popular video requests, we observe that the similarity becomes higher when closer to the downtown. Furthermore, comparing similarity according to topic types between the downtown and the suburb, we find that shows and movies exhibit the highest and lowest similarity, respectively.

V. Implications

We have studied the spatial popularity and similarity of watching videos. Based on our findings, we can infer some important implications on the cache deployment, advertisement and video recommendation.

The demand for efficient video delivery as well as the support of flexible deployment make cache an important element of future mobile network. Especially, in the city scale, ISP can deploy multiple cache appliances, with each one serving users in the deployed region. However, fundamental understandings of cache for mobile network, e.g., how cache behaves with the large-scale video contents and viewing requests, and effective cache deploying methods are still lacking. Our finding can provide a valuable guideline on strategies of cache deployment, e.g., video popularity exhibits the concentration even in small-scale region and thus cache can be deployed with a flexible scale. Further, we observe that different regions have different similarities of video popularity, which suggests that

different regions should cache different topics. In addition, the concentration and similarity of video popularity on the types of topics exist great differences between the downtown and the suburb. Thus, according to how close the region is to the downtown, cache on the contents and number of topics with specified types should be different, e.g., the topics belonging to TV series should be cached more in the downtown than in the suburb.

Online advertisement, as a valuable revenue for content providers, is often placed before the playback of videos. Our spatial video popularity characterization can assist deciding the strategies of advertisement. For example, the concentration of video popularity is higher in the regions of downtown, which suggests more advertisements should aim at these specified regions. Further, we should take advantage of the difference of the concentration on the different types between the downtown and suburb to provide the meaningful reference for the advertisement, e.g., putting more advertisements on the TV series in the regions of downtown.

Advertisement can bring substantial profits for content providers only when more views can be attracted. Thus, it is critical to recommend appropriate videos that are suitable for region-specified users, in order to bring higher investment of advertisement. Our findings can aid the design of the video recommendation system. For example, higher concentration on TV series in the downtown indicates that the corresponding videos should be recommended to the users in these regions. Furthermore, other spatial characteristics of video popularity can also be considered to efficiently implement the excellent video recommendation.

VI. Related Work

User viewing behaviors in Internet video system have been studied in many works [4], [6], [9], [16], [18]. [8] provide analysis about user behavior, video popularity and their impacts on recommendation. Abrahamsson et al. [5] study the access patterns and program popularity in a TV-on-demand system. Similarly, [11] investigate the viewing behavior and user activity pattern in the PPTV live streaming system. Li et al. [16] characterize different user behaviors of watching mobile videos from IPTV and VoD systems in comparing the 3G and WiFi access methods. In addition, Lin et al. [9] investigate peer-assisted video delivery in WiFi networks and conduct the analysis of viewing time, user population and user locality. Shafiq et al. [17] present the feature of mobile video streaming performance and model its influence on user engagement from network operators' perspective. Unlike them, we study the spatial features of video popularity and similarity from the perspective of ISP since our data contains view request of videos from multiple content providers.

There have been some studies on geographic features of watching online videos [7], [10], [12], [15]. Brodersen et al. [2] study whether YouTube videos exhibit geographic locality of YouTube videos over the world and show that most of views come from the videos in a single country. Platt et al. [14] study videos trending across several nations and make the contribution to understand the international cultural impact and potential of videos. Li et al. [3] provide an in-depth study on the geographic patterns of mobile video consumption of PPTV among different provinces within a country, and obtain the distinct geographic popularity features on popular and non-popular videos. Our work differs with these works in that we aim at exploring the spatial features of video popularity and similarity under the city scale.

VII. Conclusions

In this paper, we aim at understanding the spatial characterization of popularity and similarity of watching videos in the city scale. We carry out a comprehensive analysis of the concentration and similarity of video popularity in the spatial dimension over a large-scale video viewing logs that covers six most popular video content providers in China. Through the analysis, we obtain some interesting but important findings, which are summarized as follows:

(i) With regarding video popularity when considering different scales and region locations in a city, we find that under all four studied scales, video view requests conform with Pareto principle, and thus the concentration of video popularity exists at all these scales. Furthermore, under the same scale, the differences of concentration exist in different regions and the regions in the downtown exhibit higher concentration than ones in the suburb.

(ii) With regarding video request similarity when considering different scales and region locations in a city, we find that the the similarity of video popularity in a region varies with each other, which is especially significant under the small scale. The nearer the region is to downtown, the higher similarity there is when compared with a chosen region in the downtown.

These findings offer useful guidelines for cache deployment, advertisement and video recommendation.

References

[1] Cisco. "Cisco visual networking index: Global mobile data traffic forecast update, 2014-2019, 2015.

[2] A. Brodersen, S. Scellato, and M. Wattenhofer. "Youtube around the world: geographic popularity of videos, in *Proceedings of WWW 2012*, 241–250.

[3] Z. Li, G. Xie, J. Lin, Y. Jin, M. Kaafar, et al. "On the geographic patterns of a large-scale mobile video-on-demand system, in *Proceedings of IEEE INFOCOM 2014*, 397–405.

[4] A. Finamore, M. Mellia, M. Munafo, R. Torres. S. G. Rao, "Youtube everywhere: impact of device and infrastructure synergies on user experience, in *Proceedings of IMC 2011*, pp 345–360.

[5] H. Abrahamsson, and M. Nordmark, "Program popularity and viewer behavior in a large tv-on-demand system, in *Proceedings of the ACM IMC 2012*, pp 199–210.

[6] M. Cha, H. Kwak, P. Rodriguez, Y. Ahn, and S. Moon. "I tube, you tube, everybody tubes: analyzing the world's largest user generated content video system, in *Proceedings of the ACM IMC 2007*, 1–14.

[7] H. Yin, X. Liu, F. Qiu, N. Xia, C. Lin, H. Zhang, V. Sekar, and G. Min. "Inside the bird's nest: measurements of large-scale live vod from the 2008 olympics, in *Proceedings of the ACM IMC 2009*, 442–455.

[8] H. Yu, D. Zheng, B. Y. Zhao, and W. Zheng. "Understanding user behavior in large-scale video-on-demand systems, in *Proceedings of EuroSys 2006*, 333–344.

[9] 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, in *IEEE Communications Magazine, 2013*, 120–126.

[10] Z. Li, J. Lin, M. Akodjenou, G. Xie, M. Kaafar, Y. Jin, and G. Peng. "Watching videos from everywhere: a study of the pptv mobile vod system, in *Proceedings of the ACM IMC 2012*, 185–198.

[11] Z. Li, G. Xie, M. A. Kaafar, and K. Salamatian. "User behavior characterization of a large-scale mobile live streaming system, in *Proceedings of the WWW 2015*, 307–313.

[12] M. Cha, P. Rodriguez, J. Crowcroft, S. Moon, and X. Amatriain. "Watching television over an ip network, in *Proceedings of the ACM SIGCOMM 2008*, 71–84.

[13] M. E. J. Newman. "Power laws, pareto distributions and zipf's law, in *Contemporary Physics*, 323–351.

[14] E. L. Platt, R. Bhargava, E. Zuckerman. "The International Affiliation Network of YouTube Trends, in *Proceedings of the ICWSM 2015*.

[15] S. Scellato, C. Mascolo, M. M., and Crowcroft, J. "Track globally, deliver locally: improving content delivery networks by tracking geographic social cascades, in *Proceedings of WWW 2011*, 457–466.

[16] Y. Li, Y. Zhang, and R. Yuan. "Measurement and analysis of a large scale commercial mobile internet tv system, in *Proceedings of the ACM IMC 2011*, 209–224.

[17] Z. M. Shafiq, J. Erman, L. Ji, A. X. Liu , J. Pang, and J. Wang. "Understanding the impact of network dynamics on mobile video user engagement, in *Proceedings of the ACM Sigmetrics 2014*, 367–379.

[18] Y. Borghol, S. Ardon, N. Carlsson. "The Untold Story of the Clones: Content-agnostic Factors that Impact YouTube Video Popularity, in *Proceedings of the KDD 2012*, 1186–1194.