Shrinking VOD Traffic via Rényi-Entropic Optimal Transport

Shrinking VOD Traffic via Rényi-Entropic Optimal Transport

CHI-JEN (ROGER) LO, University of Cambridge, UK
MAHESH K. MARINA, The University of Edinburgh, UK
NISHANTH SASTRY, University of Surrey, UK
KAI XU, MIT-IBM Watson AI Lab, USA
SAEED FADAEI, University of Surrey, UK
YONG LI, Tsinghua University, China

In response to the exponential surge in Internet Video on Demand (VOD) traffic, numerous research endeavors have concentrated on optimizing and enhancing infrastructure efficiency. In contrast, this paper explores whether users' demand patterns can be shaped to reduce the pressure on infrastructure. Our main idea is to design a mechanism that alters the distribution of user requests to another distribution which is much more cache-efficient, but still remains 'close enough' (in the sense of cost) to fulfil each individual user's preference. To quantify the cache footprint of VOD traffic, we propose a novel application of Rényi entropy as its proxy, capturing the 'richness' (the number of distinct videos or cache size) and the 'evenness' (the relative popularity of video accesses) of the on-demand video distribution. We then demonstrate how to decrease this metric by formulating a problem drawing on the mathematical theory of optimal transport (OT). Additionally, we establish a key equivalence theorem: minimizing Rényi entropy corresponds to maximizing soft cache hit ratio (SCHR) — a variant of cache hit ratio allowing similarity-based video substitutions. Evaluation on a real-world, city-scale video viewing dataset reveals a remarkable 83% reduction in cache size (associated with VOD caching traffic). Crucially, in alignment with the above-mentioned equivalence theorem, our approach yields a significant uplift to SCHR, achieving close to 100%.

ACM Reference Format:
Chi-Jen (Roger) Lo, Mahesh K. Marina, Nishanth Sastry, Kai Xu, Saeed Fadaei, and Yong Li. 2024. Shrinking VOD Traffic via Rényi-Entropic Optimal Transport. *Proc. ACM Meas. Anal. Comput. Syst.* 8, 1, Article 7 (March 2024), 34 pages. https://doi.org/10.1145/3639033

1 INTRODUCTION

In recent years, users' appetite for video consumption has been growing unabated. For instance, according to Cisco's Visual Networking Index, Video on Demand (VOD) traffic accounts for 82% of the Internet traffic as of 2022 [8]. Not only does this imply an upsurging Internet traffic with larger cache sizes in content delivery networks (CDNs), but also has detrimental environmental impact. While most work focuses on mechanisms for decreasing network traffic and energy footprint of a particular video request pattern, we pose a complementary question: can we instead change the pattern of video requests in such a way that it decreases the cache (and network traffic, energy,

Authors' addresses: Chi-Jen (Roger) Lo, cjl202@cam.ac.uk, Department of Engineering, University of Cambridge, UK; Mahesh K. Marina, mahesh@ed.ac.uk, School of Informatics, The University of Edinburgh, Scotland, UK; Nishanth Sastry, n.sastry@surrey.ac.uk, School of Computer Science and Electronic Engineering, University of Surrey, UK; Kai Xu, xuk@mit.edu, MIT-IBM Watson AI Lab, Cambridge, MA, USA; Saeed Fadaei, s.fadaei@surrey.ac.uk, School of Computer Science and Electronic Engineering, University of Surrey, UK; Yong Li, liyong07@tsinghua.edu.cn, Department of Electronic Engineering, Tsinghua University, Beijing, China.

ACM 2476-1249/2024/3-ART7
https://doi.org/10.1145/3639033

etc.) footprint while keeping the users satisfied? *We present a novel solution with Rényi-entropic optimal transport, encapsulating concepts of cache-aware recommendation and soft caching.*

Cache-aware recommendation draws its inspiration from the realm of recommendation systems. Unlike conventional recommendation approaches (e.g. *content-based*, *collaborative filtering*) that emphasize only the satisfaction of an individual's preference, cache-aware video recommendation is geared towards a proactive video caching mechanism that not only caters to user preferences but prioritizes the improvement of efficiency in video delivery, leading to decreased CDN cache sizes and associated downloading traffic. Although the idea may appear simple at first glance, it is, in fact, non-trivial. *Each cache has to serve an assemblage of users with mismatched access patterns and video preferences, so it is quite challenging to come up with a cache recommendation from these ever-changing patterns, not to mention an operational method to reduce it.* One may also note that although we cannot restrict user behaviors on the 'demand-side', it remains possible to cache, instead, an alternative set of videos which still satisfy individual's preference [43]. Fortunately, from a data-driven perspective, we can condense these seemingly conflicting concerns through the language of probability and its advances, i.e., information theory and optimal transport.

We regard the video viewing events altogether as a probability distribution, where each event is a combination of features like *where*, *when*, and *which* video a user requests. A cache-friendly design favors the situation when several users are asking for the same video (during a certain period), so that multiple requests can be fulfilled by a single cached video. Given a distribution of user requests (i.e. video viewing events), we aim to find an alternate distribution to shrink down the cache size by imposing a 'clustering' effect on its feature space, i.e., generating more spatial-, temporal- or content-type overlaps. We find the mathematical theory of *optimal transport (OT)* [45] a perfect candidate to realize the change in the distribution. Initially conceived as a principled way to move a pile of sand from one place to another but with a prescribed final shape. Merging the contexts of statistics and optimization, optimal transport is popular nowadays as a means of converting a given probability distribution to another with a desired shape. Especially in (statistical) machine learning, it enables comparison of two degenerate probability distributions supported on low-dimensional manifolds in much higher-dimensional spaces [15]. Most importantly, the transport cost can be defined with a point-wise granularity. Therefore, even if what we are shifting is a 'joint' distribution of user video requests, such an ability to quantify point-to-point variation in distributions allows us to ensure a recommendation admissible for each 'individual' by setting cost constraints.

To characterize the cache size, we employ an information-theoretic statistic known as *Rényi entropy* [35]. In our setting, more overlaps in the feature space indicate a lower 'evenness' in distribution. However, this does not necessarily suggest a cache size reduction, since a steeper shape in distribution can still contain same quantity, or 'richness', of videos, as it is referred to in the ecological context. Rényi entropy generalizes several widely used metrics for capturing diversity and measures both the richness (representing the number of distinct videos) and the evenness (reflecting the popularity of each video) of an empirical distribution (of video viewing events) [18]. We formulate a problem of minimizing the Rényi entropy subject to admissibility by reshaping a given distribution of user video requests into a recommended distribution. It is further reformulated as a (convex-concave minimax) problem of Rényi-entropic optimal transport (Lemma 2). Inspired by the concept of soft caching [37], the equivalence theorem (Theorem 1), a key result of this paper, justifies that Rényi entropy minimization is equivalent to maximizing soft cache hit ratio (SCHR) — a variant of cache hit ratio (CHR) that allows user-acceptable substitutions of 'similar' videos. *As a result, a variety of separate ideas and concerns including VOD traffic shrinkage, cache size reduction, video recommendation, user admissibility, and metric learning are all brought together into a coherent whole.* To efficiently solve the above outlined problem, we propose a low-complexity algorithm termed `SteepOTVR` that scales well to cope with large datasets. Experimental results using a real

world city-scale video viewing dataset [26] show a significant shrinkage in the cache sizes, and thanks to the power of optimal transport, a vastly improved soft cache hit ratio compared to the three baseline caching algorithms considered, i.e., LRU [30], Bélády [1], and SCH [37].

In summary, we list the contributions of this paper as follows:

- First, we link the concept of 'diversity' in ecology to the size of CDN caches holding VOD content with data analytical evidence (in a spatio-temporal regime), allowing us to introduce Rényi entropy as a proxy for the CDN cache size and associated VOD traffic.
- Second, we formalize the idea of cache-efficient video recommendation into a constrained Rényi entropy minimization problem, and show how to reformulate it as a (Rényi) entropic-regularized optimal transport. *We also prove that minimizing Rényi entropy is equivalent to maximizing a variant of the familiar metric of cache hit ratio (CHR), known as the soft cache hit ratio (SCHR).*
- Finally, we design an efficient algorithm called SteepOTVR for the formulated problem, to learn a cache-friendly video distribution while maintaining individual user's satisfaction. Evaluation based on a city-scale dataset shows that our proposal cuts down Rényi entropy and cache size by 69% and 83%, whereas simultaneously increasing the SCHR from 34% to 98% — with a 2.88x gain.

The rest of the paper is structured as follows: Section 2 provides background knowledge on the mathematics applied, our dataset and the CDN model. In Section 3, we explore the patterns of video views to assess the potential of cutting down CDN cache sizes and VOD traffic via recommendation. The design of our proposed optimal transport video recommender, SteepOTVR, is presented in Section 4, with its performance evaluation in Section 5. We then give an overview of related work in Section 6 before concluding with Section 7.

2 BACKGROUND AND DATASET

2.1 Optimal Transport

Retrospecting back to 1781, mathematician Gaspard Monge formalized the problem of *optimal transportation*: an allocation problem that aims to find the minimal effort moving a pile of earth to the another place with a prescribed shape. Intuitively, these two piles of earth can be seen as two probability distributions, and an *optimal transport (OT)* is a mapping that minimizes the overall effort. Here we give the mathematical definition reformalized by Cédric Villani in [45]:

Definition 1 (The Monge's Problem on Optimal Transport). *Let $\mathcal{X}$ and $\mathcal{Y}$ be two separable metric spaces such that any probability measure on $\mathcal{X}$ or $\mathcal{Y}$ is a Radon measure and let cost $c(\cdot,\cdot) : \mathcal{X}\times\mathcal{Y} \to \mathbb{R}_0^+$ be a Borel-measurable function. Given probability measures μ on $\mathcal{X}$ and ν on $\mathcal{Y}$, Monge's formulation of the optimal transportation problem is to find a transport map $T : \mathcal{X} \to \mathcal{Y}$ that realizes the following infimum:*

$$\inf\left\{ \int_{\mathcal{X}} c(x, T(x))\, \mathrm{d}\mu(x) \;\middle|\; T_{\#}(\mu) = \nu \right\}, \tag{1}$$

where $T_{\#}(\mu)$ denotes the pushforward (i.e. image measure) of μ by T. Literally, an optimal transport is a mapping T which attains this infimum.

Monge's formulation of the transportation problem is often ill-posed. Fortunately, this could be circumvented by adopting Kantorovich's relaxation which finds a probability measure π^* upon the product space $\mathcal{X}\times\mathcal{Y}$ defined as:

$$\pi^* \triangleq \underset{\pi\in\Pi(\mu,\nu)}{\arg\inf} \left\{ \int_{\mathcal{X}\times\mathcal{Y}} c(x,y)\, \pi(\mathrm{d}x, \mathrm{d}y) \right\}, \tag{2}$$

where $\Pi(\mu,\nu)$ denotes the collection of all probability measures on $\mathcal{X}\times\mathcal{Y}$ having marginals μ on $\mathcal{X}$ and ν on $\mathcal{Y}$, also named as the *coupling* of distributions $\mathcal{X}$ and $\mathcal{Y}$. To avoid intractability, one often

resorts to the entropic regularization[1] of Kantorovich's formulation, providing *Sinkhorn distance* as:

$$W_\zeta^*(\mu,\nu) \triangleq \inf_{\pi\in\Pi(\mu,\nu)} \left\{ \int_{\mathcal{X}\times\mathcal{Y}} c(x,y)\,\mathrm{d}\pi + \zeta D_{\mathrm{KL}}(\pi||\mu\otimes\nu) \right\}, \tag{3}$$

in which $\mathrm{d}\pi \triangleq \pi(\mathrm{d}x,\mathrm{d}y)$, $\zeta > 0$ is a regularization parameter and $D_{\mathrm{KL}}(\pi||\mu\otimes\nu) \triangleq \int \log_e \frac{\mathrm{d}\pi}{\mathrm{d}(\mu\otimes\nu)}\mathrm{d}\pi$ is the *Kullback–Leibler (KL) divergence* of the coupling distribution π over *product measure* $\mu\otimes\nu$ [9].

2.2 The Shanghai Video Viewing Dataset

For the initial analysis and later evaluations, we use the video viewing dataset collected at the gateways of a major Internet service provider (ISP) in Shanghai from Nov. 1st to Dec. 31st, 2014 [26].

This dataset covers a total of 1.4 million users across Shanghai city who had pulled 200 million requests and watched around 7 million unique video (content) items, spanning the top six most popular video content providers who comprehensively support video-on-demand (VOD) service. The videos are categorized into five different genres: cartoon, movie, show, TV play and others. Each row represents a video viewing event, constituting the following features: Timestamp (T), Video id (V), Access point id (A), Video genre (G), Coordinate (C), and OS info (O). We furthermore set function $w(\cdot)$ for denoting the Views per video (id), which varies over queries of row combination. To illustrate, we show a snippet from the dataset in Table 1.

Timestamp (T)	Video ID (V)	Access Point ID (A)	Video Genre (G)	Coordinate (C)	OS Info (O)	Views per Video (w)
11-01 00:00:00	video-4	access-0	genre-2	(x_6, y_9)	iOS 10	95
11-01 00:00:01	video-3	access-0	genre-3	(x_5, y_0)	PC	12
11-01 00:00:01	video-2	access-4	genre-1	(x_3, y_2)	Android	1769
11-01 00:00:02	video-2	access-1	genre-1	(x_7, y_9)	Android	1769

Table 1. Illustrative snippet of the Shanghai video viewing dataset.

Data Validity. Even though this dataset is ten years old, it is still suitable and representative for the purposes of our study because:

- Most of the viewing patterns we rely on broadly stay the same (e.g. people still view more during evening peak hours); people still have certain genres that they prefer (e.g. some prefer drama, others prefer documentary, etc.); views are still skewed towards a smaller set of popular videos.
- Shanghai is among the largest cities in the world, so the volume of video views for Shanghai a decade back will be valid for smaller urban regions today. Thus, although the actual videos being viewed in Shanghai may have changed, we still expect to observe certain similarities in patterns.

Above all, our recommendation regime is agnostic to the change of either the features or the data, so it can be applied to any other dataset. Besides, we note that due to the rising privacy awareness in our society, it is now getting more difficult for academics to get access to any user-related data, posing difficulty for us to extend our analysis and evaluation to other newer datasets.

2.3 CDN Caching Model

In realizing our idea of reducing traffic through cache-aware video recommendations, practicalities of the VOD service delivery have to be kept in mind. We thus follow the tree structure from [3] in our network model. Since the downstream video data pass through from the providers to the caches at the network edge, cache sizes reflect the ingress traffic flow. In practice, for matching the users (video viewing events) with the local caches, we apply the modified version of *k-means clustering* in [4]. Specifically, we partition all the video viewing events spatially over the city into $k = 20$ clusters with similar quantity of events, so the clusters are assigned with corresponding caches serving a similar amount of requests, in effect uniformly distributing the viewing events

[1]Entropic optimal transport appeared first in Schrödinger's work to find the most likely particle configuration evolution [36].

across different caches (Appendix B.2). Note that our network model as outlined above is in line with commonly used edge caching system models in the literature [51].

User Feeds. The videos users receive from their feeds are dependent on the CDN cache serving the user, while individual user's feeds can still be personalized (despite its dependence on cache recommendation in our proposal). This means our design of data-driven cache-aware video recommendation is not specified to directly assign contents to the individual users, but to recommend the minimal and most likely set of videos (at proper timings) at every single local cache so that the download traffic to the CDN caches could be pared down while staying aligned with the preference and schedule of video viewing events of individual users. Under this model, we treat VOD traffic as an interchangeable concept to CDN traffic.

3 METRICS AND INSPECTIONS

In this section, we first propose the idea of applying Rényi entropy as a proxy of cache size and VOD traffic. We also assess the dataset to see if there is a potential of decreasing the traffic by changing the request patterns of users.

3.1 Rényi Diversity Index as a Proxy of Cache Size

The size of a cache is a reflection of the 'diversity' of videos it caches for the users. Some videos are extremely popular and serve multiple users, whereas some videos are of more niche interest. We can make an analogy of this distribution of videos to an ecosystem, whose diversity is a function of how they distribute (i.e. evenness) as well as the abundance or numbers of each species (i.e. richness). It is apparent that less richness leads to a decreased cache size. Yet, richness itself is not a sufficient metric to define a successful recommendation, since user preferences deviate over a wide spectrum. Otherwise, we can imagine an extreme case with only the most 'popular' video serving all the requests. It could never satisfy every user.

Index of Diversity. A variety of fields ranging from astronomy and demography to ecology have the need for measuring diversity and thus many diversity indices have been proposed. Many of them only account for the categorical diversity of the species / entities (i.e. the quantity of distinct species, sometimes known as the *richness*), but not the total variation of numbers of individuals in each category (species / entities). An ideal index should not only capture the total variation but is expected to be capable of exploring both the categorical diversity (i.e. richness) and qualitative diversity (i.e. the *evenness* of the distribution, in terms of the quantities of each species). Specifically, we list seven of the most applied diversity indices [28] in Table 2, in which only two of them can (explicitly) characterize both the richness and evenness of entities (for any distribution of interest) — *Rényi entropy* and *true diversity* (i.e. *Hill number*), the exponent of the former. All remaining metrics can be derived as a specialization of Rényi entropy (or true diversity). For instance, when $\alpha \longrightarrow 1$, Rényi entropy converges to *Shannon entropy*. Owing to its primacy, we apply Rényi entropy as a diversity index throughout this paper:

Definition 2 (Rényi entropy). *Let X represent a discrete random variable sampled from a distribution $p_i \triangleq \Pr[X = x_i], \forall 1 \leq i \leq n$. Rényi entropy defined in [35] is*

$$H_\alpha(X) \triangleq \frac{1}{1-\alpha} \log_2 \left(\sum_{i=1}^{n} p_i^\alpha \right), \tag{4}$$

parametrized by a constant $\alpha \geq 0, \alpha \neq 1$ [18]. To distinguish its usage as a diversity index, we denote $\widehat{H}_\alpha(X)$ as the Rényi entropy on an empirical distribution where p_i is taken as the proportional abundance of the ith species. Note that the index $\widehat{H}_\alpha(X)$ is a concept distinct from the estimate $H_\alpha(X)$.

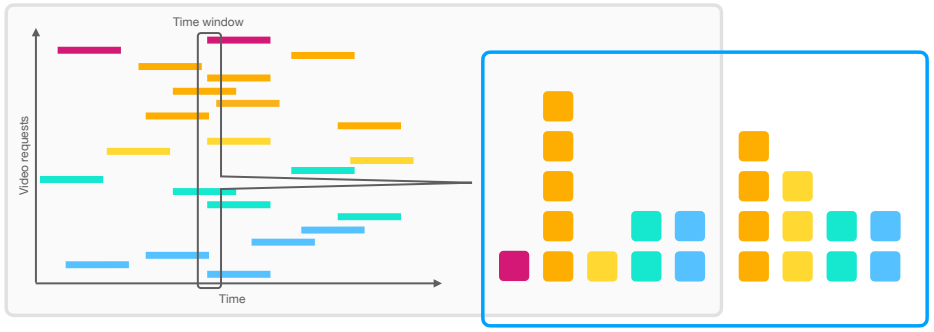


Fig. 1. The illustration has two parts. The left box (in gray) shows how the temporal video views within some certain window map to a distribution, in which each color depicts a single video; whereas the right box (in blue) is an example of two possible video distributions of 11 total views, where the left has higher richness while the right is more even. The parameter α in Rényi entropy sets a trade-off in these aspects of diversity.

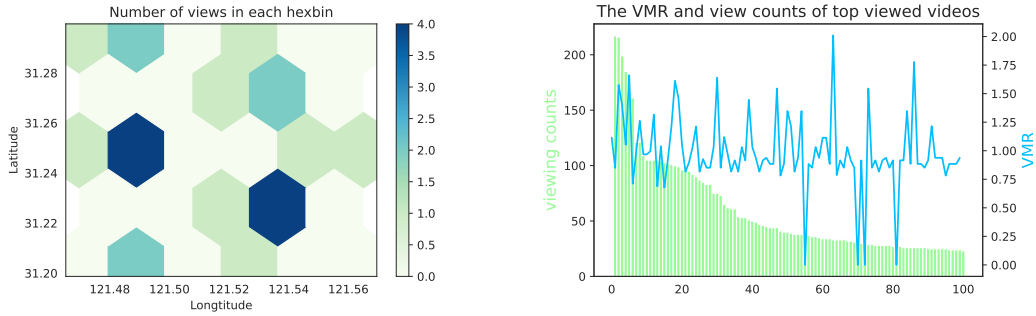
Compared to the other indices, the additional parameter α in Rényi entropy (or q in true diversity) provides an extra degree of freedom for different applications to adjust along their own criteria on diversity — to weight more on the richness or the evenness instead. If α increases, more weightage is assigned to the common species rather than the rare ones. This property allows us to emphasize the more popular videos but not the ones in the heavy-tail, and it ends up to be very useful on saving VOD traffic (as Appendix B.4 evinces). One shall also notice that Rényi entropy decreases monotonically when α increases, stretching out a full spectrum of entropies.

Richness	$R = 1/{}^0H$
Shannon Entropy	${}^1H = \log({}^1D)$
Simpson Dominance	$\lambda = 1/{}^2H$
Gini–Simpson Index	$\overline{\lambda} = 1 - \lambda$
Berger–Parker Index	$B = 1/{}^\infty H$
True Diversity	${}^qD = \left(\sum_{i=1}^R p_i^q\right)^{1/(1-q)}$
Rényi Entropy	${}^\alpha H = \log({}^qD)\vert_{q=\alpha}$

Table 2. Seven representative indices of diversity. In particular, ${}^1D = 1/\prod_{i=1}^R p_i^{p_i} = \exp\left(-\sum_{i=1}^R p_i \log p_i\right)$.

Inspired by [24], we give an intuition using the example in Fig. 1. The graph on the left depicts video requests for different videos (with each video marked by a unique color). The *distribution* of requests in the small time window marked in black can be depicted as blocks as shown within the intersection of the blue and gray box. This shows each video access as a stacked block. There are 11 total requests towards five different videos, with one (orange video) being extremely popular and obtaining five views, whereas others such as the pink and yellow video only receive one view each. If we could somehow change the actual distribution of requests (on the left of the blue box) to the distribution of requests on the right, the cache is still able to serve 11 accesses but with a smaller cache size of four videos. Nonetheless, if we calculate the Rényi entropy of both cases, although the right case has less 'richness', it does not guarantee a lower entropy. It depends on how α is parametrized. In our case, $\alpha > 1$ is preferable due to our concentration on cache sizes, while coincidentally it convexifies our formulation.

We formalize the above intuition in Section 4 by proposing a minimax problem that 'learns' a video distance metric to change the distribution of accesses so as to decrease cache size and traffic. We also present how minimizing Rényi entropy is equivalent to maximizing the soft cache hit ratio (SCHR). They all come after the next subsection, where we examine whether there are spatial and temporal viewing patterns that can be learned and exploited to reduce CDN cache size.



(a) An illustration of the map divided into hexagons. Each color represents a number which is the quantity of samples within their corresponding hexbin.

(b) The VMR and view counts of the top 100 viewed video IDs around 1 indicates the randomness of the distribution.

Fig. 2. The spatial video viewing patterns.

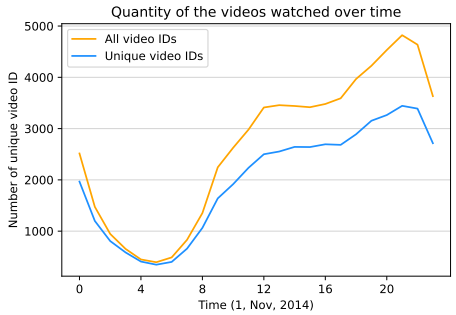
3.2 Analysis of Video Viewing Patterns

In this part we inspect and characterize the video viewing patterns of the Shanghai dataset [26], targeted to identify whether or not we can benefit from potential change of video demands.

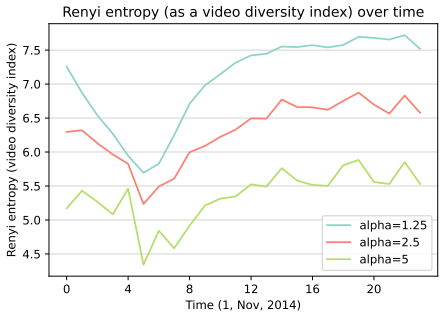
3.2.1 The Spatial Video Viewing Patterns. The *Variance-to-Mean Ratio (VMR)* (i.e. *Fano factor*) [11] is a commonly used index of dispersion which divulges the distribution pattern of samples. Its calculation is quite easy. First, disintegrate the spatial samples on the map into hexagons like Fig 2a. Knowing the quantity of samples within each hexbin, we obtain another empirical distribution whose variance over mean is the ratio we aim. In particular, a degenerate distribution leads to a VMR close to 0; a Poisson (i.e. spatially random) distribution leads to a VMR around 1; and a clustered / clumped distribution leads to a relatively large VMR. Despite its simplicity, VMR still provides strong indications to the distribution patterns of spatial point processes. In Fig. 2b we show with the blue signal the VMR of the top 100 viewed videos as well as their total views with the green bars. [2] It could be inferred that spatial distributions of these 100 videos are pretty much random as their VMR, despite oscillating back and forth, always reverts to 1. Besides, it is sufficient to conclude that we have a Poisson distribution over any combination of video IDs by the additive property of *Poisson (point) processes*.

The spatial distribution indicator VMR reveals that each individual video distributes randomly over the space when zooming into relatively thin slices of time periods. Since there is no spatial affinity of videos to specialized regions, there still exists some room to cut the traffic down by increasing correlations at the edge. Clearly, by putting the same video closer to each others, we can expect a shrinkage to the overall quantity of videos required at each cache owing to the clumped and integrated users' interests. Moreover, if a particular video is not available in its local cache, more spatial correlation implies higher likelihood of an available substitution in a neighboring cache. Alternately, we obtain better cache reuse in this manner if redundant copies of the same content are dropped by the nearby caches. To be more precise, Rényi entropy is significantly decreased when we reshuffle the spatial distribution of the videos into clusters. Such a decrease of (empirical)

[2]In this case it is still effective even if we only assess the top viewed videos. Because of the heavy-tailed nature of video views, every video in the tail is unique and has only been accessed once. Thus it becomes meaningless for the application of VMR method even if we could still obtain a value.



(a) The number of video IDs and unique ones over time. It is interesting to observe that the videos watched are mostly unique during off-peak hours. All the remaining dates come up with patterns alike.



(b) The empirical Rényi entropies of the sets of videos watched over time, under different values of α. Note that for a distribution $X \sim \boldsymbol{p}$, its empirical Rényi entropy $\widehat{H}_\alpha(X)$ strictly decreases in the value of α.

Fig. 3. The temporal video viewing patterns.

Rényi entropies over space links to smaller cache sizes, and is finally led to our goal of cutting down the VOD traffic.

3.2.2 The Temporal Video Viewing Patterns. It is well known that user accesses are higher during certain times of day [21]. The key question we want to ask here is: If the access time of users' preferable videos are shifted within a bearable extent after recommending, is there a potential of traffic decrease? To start with, we presume that each video watching event consumes the same amount of time, T, thereby the period of each event $\boldsymbol{x}_i$ can be represented by a time interval $[t_i, t_i + T]$. To prevent memory blow-up of a single cache serving multiple access points, whenever a video is downloaded / cached, the total amount of time they remain in cache (i.e. time-to-live, TTL) are identical. It is given under these mild assumptions that the cache size is proportional to the VOD traffic. Similar to the case above, Rényi entropy, while seizing the temporal video diversity, can also be seen as a proxy of the cache size. In accordance, the shrinkage of Rényi entropy over time indicates the decrease of the cache sizes and traffic temporally.

In Fig. 3a, the gold curve presents the time series of the total amount of videos requested while the skyblue curve considers only unique video IDs. The difference between these curves implies a duplication of videos fetched, which enhances during the peak hours but shrinks during off-peaks. The swelling gap along the mound indicates that there might be several popular videos that have been watched considering the long tail nature of the network traffic. Referring to the *Dirichlet's Box / Drawer Principle* (i.e. pigeonhole principle), we find that video tastes tend to be more niche, unique and dissimilar from other viewers during the midnight and dawn. Such a vanishing gap may imply that users tend to watch videos by their own choices beyond popular videos, e.g., videos not highlighted on the front page. Comparison between Fig. 3a and Fig. 3b illustrates that Rényi entropy is a good proxy of the cache sizes and, thus, its VOD caching traffic. It is evident that there are still a high proportion of users who ask for niche content (i.e. unique video IDs). Enhanced cacheability and traffic savings are attainable if we can further increase duplicate viewing counts.

3.3 Towards VOD Traffic Shrinkage — An Operational Perspective

We now distil the patterns from the previous subsection and develop schemes with the potential of reducing VOD traffic through cache-aware video recommendation.

3.3.1 Demand-Side Management and Its Challenges. The goal of our work is to make user requests more cache-efficient and environmental-friendly. Originally proposed as a mechanism to lower energy consumption in grids, demand-side management (DSM) has already been widely applied in the energy industry for decades. By intentionally shaping user behaviors, managing authorities can achieve certain goals. In particular, Tyson, Sastry, Mortier, and Feamster [43] were the first to suggest the idea of space-, time- and content-shifting, which they name *Staggercast*, aiming at decreasing the Internet traffic by recommendation. As a straw man proposal, it does not provide any metric or operational methodology on how to execute those shifts thereby the 'managed' video demand pattern can fit individual interests. However, the intention to directly change the demand-side distribution is also considered as the major rebound of nudging user demand. Individuals may not feel comfortable with the changes unless they are 'small' enough (kept within some well-defined metric). Therefore, even if the analyses in Section 3.2 validate the efficacy of doing space-shifting and time-shifting to individual's access pattern, we tend not to keep this perspective for our design, but to look from a caching point of view.

One of our major contributions is the proposal of a mechanism on cache-aware video recommendation that resolves the previous issues. In fact, we can regard space- and time-shifts as side effects of our caching scheme under a larger (spacetime) scale. *While holistically shrinking the overall cache size by the means of demand-side recommendation, changes are bounded from above through a learned cost metric in such way that every preference shift is kept admissible from the perspective of individuals. A successful recommendation is, namely, the one giving a user-admissible set of videos that reduces the cache size (and VOD traffic).* We find the theory of optimal transport is a perfect candidate for formalizing the problem mathematically. Section 4.3 gives further analyses.

3.3.2 Shrinking VOD Traffic through Its Proxy. From the recommendation point of view, one can release some workload to the CDN caches by creating more video overlaps either spatially or temporally and by leading users to watch those more popular videos apart from their original choices. In other words, the increase of such overlaps indicates not only users' share of interests but also the time and location they send a video request. *Nonetheless, beyond increasing the overlaps themselves (which reduces richness), how they distribute, uniform or uneven (comes in a 'steep' shape after sorting), should also be part of our concern. An alternative distribution, even if it had a much smaller richness (i.e. smaller cache footprint), would never be accepted if they are not acceptable by individual users.* We cannot expect an extreme case that tends to satisfy all users by recommending only a few popular videos.

An instant observation referring to Section 3.1 is that 'diversity', capable of quantifying richness and evenness of a distribution, is the key to associate those overlaps with the cache size. In particular, Rényi entropy, by quantifying the scales of richness and evenness of the distribution, endows us the ability to capture the diversity of video viewing events. One can imagine that a higher diversity indicates a more 'disperse' distribution of the events in the higher-dimensional featural (Radon) space, whereas such a 'dispersion' would inevitably consume more CDN resources on caching the videos either from the perspective of space, time, or content; and vice versa. In this vein, the cache size should correlate positively to Rényi entropy under the same distribution. *On top of that, if we slightly relax the condition of video overlaps so that videos with similar content are interchangeable subject to some user acceptable threshold, then the cache sizes could be further suppressed by clustering 'similar' videos closer in spacetime as long as a 'similarity' metric is properly set up.* We discuss this in the next passage, with detailed design rolled out in Section 4.1.

3.3.3 Constrained Clustering in the Feature Space. In Section 3.2 we noted that cache size can be decreased by creating opportunities for time-shifting (e.g. delay) or space-shifting (e.g. redirecting

the user to a neighboring cache). While, rather than consider each feature such as space and time individually, we can benefit by taking a holistic view to address several issues:

- *Side-information.* In prior inspections, only the labels of videos (VIDEO ID) are reflected considering space- and time-shifts. Nonetheless, in real recommendation systems, it would be better to generate video embedding (in a finer sense) that involves more contextual information.
- *Distance metric.* This is a natural issue that arises in content-shifts. Since our dataset only provides a few features directly linked to the content of videos (VIDEO GENRE), we have to figure out some way to recover these distances. Moreover, the other features providing contextual information should also be inlet as a part of the bigger picture of distance metric generation (as in Section 4.2).
- *Featural correlations.* Regarding each video request (i.e. event) as a featural combination, a reasonable way to decrease traffic demand is to cluster these events by their features. This counterexample indicates the need of a comprehensive consideration of distinct features: Suppose someone is imposed by recommendation a swap of two video IDs over the time arrow (i.e. temporal), then such an operation will inevitably lead to the shift of videos (i.e. content), when we look at the isolated spatial snapshots at these timestamps of interchange. Thus, we can expect some changes on the spatial video diversities at these timestamps (even if those are slight). This tells the reason why we should not cope with these features independently.

These all suggest that the content recommendation point of view is more preferable. No matter how the access pattern is distributed spatially or temporally, one focuses on finding the 'optimal' set of videos to be cached. We once again give an intuition through Fig. 1. If the pink video is very 'similar' to the yellow video (according to some notion of similarity), one can simply *recommend* the yellow video to the pink video user instead, in such way that the richness of videos reduces by 1 (from 5 to 4). To address content-shifts as well as the listed issues, we present in Section 4 the train of thought behind our coherent design of the video recommender, followed by its novel realizations.

4 THE OPTIMAL TRANSPORT VIDEO RECOMMENDER

In this section, we first bring forward the provisions towards a successful recommendation, then dive into the technical details of the optimal transport video recommender.

4.1 Constrained Rényi Entropy Minimization

We pursue in this section an operational method that leads to a successful recommendation on arbitrary dataframe $\boldsymbol{X} \triangleq \{\boldsymbol{x}_i\}_{i=1}^{m_x}$, where $\boldsymbol{x}_i \in \mathbb{R}^n$ is the ith row of $\boldsymbol{X}$ recording the information of a viewing event like: when, where, and which video the user watches, drawn from some featural *Radon space* $\mathcal{X} \subset \mathbb{R}^n$. A recommendation imposes an one-to-one shift from each $\boldsymbol{x}_i$ to some corresponding $\boldsymbol{y}_j \in \mathbb{R}^n$ drawn from another feature space $\mathcal{Y} \subset \mathbb{R}^n$, and thus forms a transport map between $\boldsymbol{X}$ and $\boldsymbol{Y} \triangleq \{\boldsymbol{y}_j\}_{j=1}^{m_y}$. Generally, we assume the total choices of video viewing events (i.e. number of rows) remains unchanged: $m_x = m_y = m$. Since we can simply assign $\boldsymbol{q}_j = 0$ to those $\boldsymbol{y}_j$ not considered, this covers the situation of all $m_y \leq m_x$. Further, we assign empirical distributions $\mu \triangleq \sum_{i=1}^m p_i \delta_{\boldsymbol{x}_i}$ and $\nu \triangleq \sum_{j=1}^m q_j \delta_{\boldsymbol{y}_j}$ uniquely over spaces $\mathcal{X}$ and $\mathcal{Y}$ with probability simplexes $\boldsymbol{p} \in \mathbb{R}^m$, $\boldsymbol{q} \in \mathbb{R}^m$ on their corresponding *Dirac delta measures* (Appendix A.1).

Unlike common formulations of deterministic optimizations, our probabilistic framework embraces several advantages. First, the randomness of data / events is carried by the nature of probability. Second, it paves a way for the adoption of a wide range of statistical metrics. To take advantage of this, we establish the following program of constrained *Rényi entropy minimization*:

$$\begin{aligned} &\text{Minimize} \quad \widehat{H}_\alpha(Y) \\ &\text{subject to} \quad (\boldsymbol{x}_i, \boldsymbol{y}_j) \in \mathcal{D}, \ \ \forall 1 \leq i, \ j \leq m, \end{aligned} \tag{5}$$

with $Y \sim \nu$ and $\mathcal{D} \triangleq \{(\boldsymbol{x}_i, \boldsymbol{y}_j) \mid d_{S^\dagger}(\boldsymbol{x}_i, \boldsymbol{y}_j) \leq 1\}$ as a compact convex set in vector space identifying the ellipsoidal regions of acceptable shifts when recommendation takes place. We specify here the shifting cost by setting a function

$$c(\boldsymbol{x}_i, \boldsymbol{y}_j) \triangleq d_S(\boldsymbol{x}_i, \boldsymbol{y}_j) \triangleq \sqrt{(\boldsymbol{x}_i - \boldsymbol{y}_j)^T S(\boldsymbol{x}_i - \boldsymbol{y}_j)} \in [0, 1], \tag{6}$$

constituting the generalized *Mahalanobis distance* parametrized by some positive semidefinite matrix S (i.e. $S \succeq 0$). Nevertheless, one may still identify a deficiency in (5) of lacking an assured cost / distance metric that corresponds to a specific $S^\dagger \succeq 0$. To overcome this, we will give in the next subsection the reformulation of (5) into a minimax optimal transport problem (Lemma 2) involving the learning of a distance metric.

The objective of Rényi entropy $\widehat{H}_\alpha(Y)$ in (5) has two effects: First, minimizing the objective decreases the video diversity acting as a proxy of the cache sizes. Thus, we can expect a saving on the overall VOD traffic by solving (5). Second, when searching for the transport map, like the entropic regularization in [9], Rényi entropy also serves as a relaxation to circumvent intractability. In fact, when taking $\alpha \longrightarrow 1$, our generalized regularizer degenerates into the original entropic regularizer in terms of Shannon.

4.2 Problem Reformulation — Optimal Transport Video Recommendation

Here we demonstrate how we reformulate the Rényi entropy minimization problem exploiting the theoretical framework of optimal transport. Such a reformulation endows an extra ability to learn a distance metric for the video embeddings. It enables us to find an 'optimal' precision matrix $S^\dagger \succeq 0$ and its associated cost function $c : \mathcal{X} \times \mathcal{Y} \longrightarrow \mathbb{R}_0^+$, compensating aforesaid deficiency.

4.2.1 Learning the Cost of Event-Shifts. No matter which *modus operandi* one resorts to, in the mission of recommendation, knowing the relations among videos (or events in our context) is of the utmost importance. Inspired by Section 3, we presume the videos a viewer prefers correlate to the category they belong to (Video genre), place where they are requested (Coordinate), time when they are watched (Timestamp), as well as their streaming devices (OS info).

In an attempt to learn the distances between each pair of video viewing events from the dataset, we introduce a *machine learning (ML)* discipline named *distance metric learning* in Xing, Ng, Jordan, and Russell [46]. Let $\boldsymbol{X} \triangleq \{\boldsymbol{x}_i\}_{i=1}^m$ be (an extraction of) the dataset that provides side-information to an additional column named Event so long as $\boldsymbol{x}_i$ undertakes the information of the ith event. Since we can only query side-information, the distance between pairing $\boldsymbol{x}_i$ and $\boldsymbol{y}_j$ can only be learned in a *weakly-supervised* (while some say unsupervised) sense. By setting up a distance metric of the form $d_\Omega(\boldsymbol{x}_i, \boldsymbol{y}_j) \triangleq \sqrt{(\boldsymbol{x}_i - \boldsymbol{y}_j)^T \Omega (\boldsymbol{x}_i - \boldsymbol{y}_j)}$ that parameterizes a family of Mahalanobis distances with some $\Omega \succeq 0$, Xing *et al.* [46] designed the following learning problem:

$$\begin{aligned} \text{Minimize} \quad & \sum_{\mathcal{L}^+} (\boldsymbol{x}_i - \boldsymbol{y}_j)^T \Omega (\boldsymbol{x}_i - \boldsymbol{y}_j) \\ \text{subject to} \quad & \sum_{\mathcal{L}^-} \sqrt{(\boldsymbol{x}_i - \boldsymbol{y}_j)^T \Omega (\boldsymbol{x}_i - \boldsymbol{y}_j)} \geq 1, \\ & \Omega \succeq 0, \end{aligned} \tag{7}$$

where $\mathcal{L}^+$ stands for the set of similar pairs of points and $\mathcal{L}^-$ stands for the set of dissimilar pairs of points [46]. More on this method and the selection of $\mathcal{L}^+$ are given in Appendix B.4.

In correspondence, even though the 'learned' $\Omega^\dagger \succeq 0$ and its corresponding cost function $c^\dagger$ are not given *a priori* in definitions (5) and (6), from where relevant we can still generalize the problem itself by further embedding the idea of distance metric learning into its context:

Lemma 1 (Learning the Shifting Cost). *Wielding the idea of distance metric learning with side-information in (7) [46], the shifting cost can be learned by solving the following concave program:*

$$\begin{aligned} \textit{Maximize} \quad & \sum_{i,j} \gamma_{ij} \sqrt{(\boldsymbol{x}_i - \boldsymbol{y}_j)^T \boldsymbol{S} (\boldsymbol{x}_i - \boldsymbol{y}_j)} \\ \textit{subject to} \quad & \sum_{\mathcal{L}^+} (\boldsymbol{x}_i - \boldsymbol{y}_j)^T \boldsymbol{S} (\boldsymbol{x}_i - \boldsymbol{y}_j) \leq 1, \\ & \boldsymbol{S} \succeq 0. \end{aligned} \tag{8}$$

Appendix A.2 proves the equivalence between (8) and (7) [46] on learning a distance metric in a weakly-supervised regime when $\mathcal{L}^-$ includes all possible pairings. One should also note that such an extension of (cost) learning problem is not only substantial but even inevitable in miscellaneous problems, influencing both the geometric structure as well as the context the problem convoys [10].

4.2.2 Seizing the Total Variation via Optimal Transport. Optimal transport has various advantages. Unlike divergences (i.e. Kullback–Leibler, Jensen–Shannon, and their generalization of f-divergence) or distances (i.e. total variation distance) which only measure the 'overlaps' between distributions, optimal transport has the ability to handle measures with non-overlapping supports, herewith the overall transportation cost captures the operational 'displacement' of one distribution to another. Such an ability to compare two degenerate probability distributions supported on low-dimensional manifolds in higher-dimensional spaces is a crucial factor in machine learning [15]. It is therefore no surprise that optimal transport metrics have emerged as a promising toolkit for data-intensive scientific research nowadays. Most importantly, it confers an extra degree of freedom by its cost function that can be tailored for the problems of interest.

Intending to leverage these advantages, it is natural to address our recommendation problem (5) based on the theory of optimal transport. *Novel application of optimal transport encapsulates the point-wise shifting cost all potential video substitutions, which enables us to fulfil individual user's preference even when we are shifting a joint distribution of all users holistically.* To start with, we show that (5) can be transformed into a convex-concave minimax program embedded with the metric learning of shifting cost (Lemma 1).

Lemma 2 (Rényi-entropic Minimax Optimal Transport). *Define* $\boldsymbol{C} \triangleq \sum_{i=1}^{m} \sum_{j=1}^{m} c(\boldsymbol{x}_i, \boldsymbol{y}_j) \boldsymbol{e}_i \boldsymbol{e}_j^T$ *as the cost matrix composed of concave differentiable cost function* $c : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}_0^+$. *We can then reformulate (5) into the form of (entropic-regularized) minimax optimal transport:*

$$\min_{\boldsymbol{\gamma} \in \Gamma(\mu,\nu)} \max_{\boldsymbol{S} \in \mathcal{S}^+} \left\{ \langle \boldsymbol{C}, \boldsymbol{\gamma} \rangle_F + \zeta \widehat{H}_\alpha(\boldsymbol{Y}) \right\}, \tag{9}$$

where $\mathcal{S}^+ \triangleq \{\boldsymbol{S} \succeq 0 \mid \sum_{\mathcal{L}^+} (\boldsymbol{x}_i - \boldsymbol{y}_j)^T \boldsymbol{S} (\boldsymbol{x}_i - \boldsymbol{y}_j) \leq 1\}$, $\zeta > 0$ *is a parameter, and* $\mathcal{L}^+$ *is the selected set of similar pairs of events (that makes* $\mathcal{S}^+$ *compact). In the case of* $\alpha > 1$, (9) *reduces to an equivalent convex-concave minimax optimization problem as follows:*

$$\min_{\boldsymbol{\gamma} \in \Gamma(\mu,\nu)} \left\{ \max_{\boldsymbol{S} \in \mathcal{S}^+} \langle \boldsymbol{C}, \boldsymbol{\gamma} \rangle_F + \frac{\zeta \alpha}{1 - \alpha} \log_2 \left(\|\boldsymbol{\gamma}\|_{1,\alpha} \right) \right\}, \tag{10}$$

in which $\| \cdot \|_{1,\alpha}$ *follows the definition of entry-wise* $L_{p,q}$*-norm of a matrix.*

Proof. See Appendix A.3. □

To decrease the cache sizes, we aim at a new video distribution with a smaller item cardinality that remains admissible across all users, by emphasizing 'richness' over 'evenness'. This implies we are only interested in particular the case of larger α, so we can plausibly take $\alpha > 1$ for simplicity.

4.3 Soft Cache Hit Ratio (SCHR)

The major purpose of this section is to show that our proposed method, beyond reducing cache sizes and VOD traffic, could also enhances the performance of video recommendation, with the improvement built upon the concept of 'soft caching' [37].

4.3.1 Extending Cache Hit Ratio to 'Soft' Cache Hit Ratio (SCHR). Most work on CDN caching measures performance using the *cache hit ratio (CHR)*:

Definition 3 (Cache Hit Ratio). *The CHR widely used in all contexts of caching is:*

$$\chi_{\mathrm{CH}} \triangleq \frac{\#\text{ of cache hits}}{\#\text{ of cache hits} + \#\text{ of cache misses}} = \frac{\#\text{ of cache hits}}{\#\text{ of requests}} \in [0,1], \tag{11}$$

which intuitively defines the proportion that the requested content are cached among all pull requests.

Soft cache hit ratio (SCHR), as a natural extension, relaxes χ_{CH} by further accepting other user-admissible recommendations [37]. That is to say, for an individual request, as long as it is determined as 'acceptable' by the user, soft caching allows recommendation from an increased set of items, rather than a single specified item being directly sent to the user. Exploiting this property, we thus provide a mechanism to jointly bring down the cache size as well as its corresponding VOD traffic by user-admissible video recommendations.

The SCHR aligned with our context is written as follows:

Definition 4 (Soft Cache Hit Ratio). *The SCHR translated into our context becomes*

$$\chi_{\mathrm{SCH}} \triangleq \sum_{i,j} \gamma_{ij} \mathbb{1}[c^{\dagger}_{ij} \leq 1] \in [0,1], \tag{12}$$

where $c^{\dagger}_{ij}$ denotes the price of an event shift from $\boldsymbol{x}_i$ to $\boldsymbol{y}_j$, the (i,j)-element of learned cost matrix $\boldsymbol{C}^{\dagger}$ from (10). The coupling γ_{ij} is the probability of recommending event $\boldsymbol{y}_j$ as a substitution of $\boldsymbol{x}_i$ and $\mathbb{1}[\cdot]$ is the indicator function which equals to 1 when condition holds and 0 otherwise.

Therefore, we count it a 'soft' cache hit whenever we find one or more cached videos $j \in [n]$ that can be recommended as a substitution of the original request to the ith video, such that the cost in (10) is acceptable to the user. If more than one video $j \in [n]$ can be recommended to replace the ith video with acceptable costs, then the coupling measure γ_{ij} captures the (marginal) likelihood of assigning alternative videos, attaining a soft cache hit if the mapping does not exceed the similarity metric learned in the form of transport cost. This generalizes the traditional notion of a cache hit, because a request for one item can be satisfied by serving an alternative video 'acceptably close'. Indeed, we may even fulfil a normal 'hard' cache hit under this definition, if a request for the ith video ends up being satisfied by serving the exact same video.

Thanks to well-defined point-wise transport cost, a formulation through optimal transport is capable of coping with user satisfaction in a granular sense in the soft caching regime.

4.3.2 Equivalence between Rényi Entropy Minimization and SCHR Maximization. We target the shrinkage of VOD downloading traffic by letting Rényi entropy as proxy of cache sizes. Nonetheless, in the context of CDN cache placement, it is not the entropies but cache hit ratios that are used traditionally as the metric of performance. To justify our choice, below we build an analytical link between the optimization of SCHR and the proposed use of Rényi entropy:

Theorem 1 (Equivalence). *In the case of $\alpha > 1$, Rényi entropy minimization admits an equivalent program on maximizing SCHR. That is, there exists an optimal coupling satisfying*

$$\boldsymbol{\gamma}^{*} = \operatorname*{arg\,min}_{\boldsymbol{\gamma} \in \Gamma(\mu,\nu)} \widehat{H}_{\alpha}(\boldsymbol{Y}) = \operatorname*{arg\,max}_{\boldsymbol{\gamma} \in \Gamma(\mu,\nu)} \chi_{\mathrm{SCH}}. \tag{13}$$

Proof. See Appendix A.4. □

Given this equivalence, we adopt SCHR as a metric to evaluate our recommender in Section 5.

4.4 Solution Algorithm

This subsection gives a quadratic-time algorithm as the core of our proposed recommender. Recall in Section 2.1 we talked about the entropic regularization of Kantorovich's formulation as a computationally tractable approximation. Algorithm for such regularized problem is already canonicalized and well developed in [9]. However, we advance the state-of-the-art. Due to the probabilistic nature in our formulation, our (regularized) optimal transport program (10) maintains equivalent to (5), rather than just a result of an ordinary Lagrange relaxation. (See Appendix A.2 for thorough descriptions.) On top of that, although the convexity of our problem guarantees its polynomial-time solvability, we seek an improvement in its complexity.

4.4.1 Algorithm. We present `SteepOTVR`, a 'steepest ascent' algorithm that solves the optimal transport video recommendation problem by 'steeping the distributions'. Referring to the ideas of *Frank–Wolfe algorithm* (i.e. *conditional gradient*), our proposed algorithm achieves a nearly quadratic-time solvability. As a projection-free optimization method utilizing the idea of first-order Taylor expansion, Frank–Wolfe updates are solved directly over the constrained sets parametrized by iterative constants $\psi_k \in (0,1)$, $k \in \mathbb{Z}$ by:

$$\boldsymbol{S}^{k+1} = (1-\psi_k)\boldsymbol{S}^k + \psi_k \operatorname*{arg\,max}_{\widetilde{\boldsymbol{S}}\in\mathcal{S}^+} \left\langle \nabla_{\boldsymbol{S}} J(\boldsymbol{\gamma}^k, \boldsymbol{S}^k), \widetilde{\boldsymbol{S}} \right\rangle_F, \tag{14}$$

$$\boldsymbol{\gamma}^{k+1} = (1-\psi_k)\boldsymbol{\gamma}^k + \psi_k \operatorname*{arg\,min}_{\widetilde{\boldsymbol{\gamma}}\in\Gamma(\mu,\nu)} \left\langle \nabla_{\boldsymbol{\gamma}} J(\boldsymbol{\gamma}^k, \boldsymbol{S}^k), \widetilde{\boldsymbol{\gamma}} \right\rangle_F. \tag{15}$$

Like these update equations demonstrate, in each step we ought to search for some elements in the constrained sets orthogonal to the gradients in the sense of Frobenius inner product.

In the following, we first show by Lemma 3 the calculation of the gradients respecting to our problem, then derive the corresponding optimization oracles with low computational complexities.

Lemma 3 (Gradients). *Let $J(\boldsymbol{\gamma}, \boldsymbol{S})$ be the objective of problem (10) that we want to solve. Then the respective gradients of J over $\boldsymbol{S} \succ 0$ and $\boldsymbol{\gamma}$ can be calculated with equations*

$$\nabla_{\boldsymbol{S}} J(\boldsymbol{\gamma}, \boldsymbol{S}) = \sum_{\boldsymbol{x}_i \neq \boldsymbol{y}_j} \frac{\gamma_{ij}(\boldsymbol{x}_i - \boldsymbol{y}_j)(\boldsymbol{x}_i - \boldsymbol{y}_j)^T}{\sqrt{(\boldsymbol{x}_i - \boldsymbol{y}_j)^T \boldsymbol{S} (\boldsymbol{x}_i - \boldsymbol{y}_j)}}, \tag{16}$$

$$\nabla_{\boldsymbol{\gamma}} J(\boldsymbol{\gamma}, \boldsymbol{S}) = \boldsymbol{C} + \frac{\zeta\alpha}{1-\alpha} \frac{\left[\sum_{i,j} \|\boldsymbol{e}_j^T \boldsymbol{\gamma}\|_1 \boldsymbol{e}_i \boldsymbol{e}_j^T\right]^{\circ(\alpha-1)}}{\log_e 2 \times \|\boldsymbol{\gamma}\|_{1,\alpha}^{\alpha}}. \tag{17}$$

Proof. See Appendix A.5. □

4.4.2 Optimization Oracles. Here we aim to find low-complexity oracles for conditional gradients. Starting with matrix relation $\langle \boldsymbol{A}, \boldsymbol{B} \rangle_F = \operatorname{tr}(\boldsymbol{A}^T \boldsymbol{B})$ we can write

$$\boldsymbol{S}^\star = \operatorname*{arg\,max}_{\widetilde{\boldsymbol{S}}\in\mathcal{S}^+} \left\langle \nabla_{\boldsymbol{S}} J(\boldsymbol{\gamma}, \boldsymbol{S}), \widetilde{\boldsymbol{S}} \right\rangle_F = \operatorname*{arg\,max}_{\widetilde{\boldsymbol{S}}\in\mathcal{S}^+} \operatorname{tr}\left[\nabla_{\boldsymbol{S}} J(\boldsymbol{\gamma}, \boldsymbol{S})^T \widetilde{\boldsymbol{S}}\right], \tag{18}$$

$$\boldsymbol{\gamma}^\star = \operatorname*{arg\,min}_{\widetilde{\boldsymbol{\gamma}}\in\Gamma(\mu,\nu)} \left\langle \nabla_{\boldsymbol{\gamma}} J(\boldsymbol{\gamma}, \boldsymbol{S}), \widetilde{\boldsymbol{\gamma}} \right\rangle_F = \operatorname*{arg\,min}_{\widetilde{\boldsymbol{\gamma}}\in\Gamma(\mu,\nu)} \operatorname{tr}\left[\nabla_{\boldsymbol{\gamma}} J(\boldsymbol{\gamma}, \boldsymbol{S})^T \widetilde{\boldsymbol{\gamma}}\right]. \tag{19}$$

Interestingly, it can be observed that the extreme point chosen by Frank–Wolfe, while targeting the operations of steeping long tail video distributions, coincides with the direction of steepest ascent / descent, which altogether inspire the naming of the `SteepOTVR` algorithm.

Algorithm 1: `Minimax Frank–Wolfe (Oracle I + Oracle II)`

Data: $\boldsymbol{X} = \{\boldsymbol{x}_i\}_{i=1}^m, \boldsymbol{Y} = \{\boldsymbol{y}_j\}_{j=1}^m, \boldsymbol{\gamma}^0 \in \mathcal{B}, \boldsymbol{S}^0 \succ 0, \alpha > 1, \zeta \gg \epsilon > 0$
Result: $\boldsymbol{\gamma}^k \in \mathcal{B}, \boldsymbol{S}^k \succ 0$
while $k \geq 0$ **do**
  $\nabla_S J^k \leftarrow \sum_{\boldsymbol{x}_i \neq \boldsymbol{y}_j} \gamma_{ij}^k (\boldsymbol{x}_i - \boldsymbol{y}_j)(\boldsymbol{x}_i - \boldsymbol{y}_j)^T / \sqrt{(\boldsymbol{x}_i - \boldsymbol{y}_j)^T \boldsymbol{S}^k (\boldsymbol{x}_i - \boldsymbol{y}_j)}$; // $\nabla_S J$ and $\nabla_{\boldsymbol{\gamma}} J$
  $\nabla_{\boldsymbol{\gamma}} J^k \leftarrow \sum_{i,j} \gamma_{ij}^k \sqrt{(\boldsymbol{x}_i - \boldsymbol{y}_j)^T \boldsymbol{S}^k (\boldsymbol{x}_i - \boldsymbol{y}_j)} \boldsymbol{e}_i \boldsymbol{e}_j^T + \zeta \frac{\log_2 e}{1-\alpha} \left(\alpha \left[\sum_{i,j} \|\boldsymbol{e}_j^T \boldsymbol{\gamma}\|_1 \boldsymbol{e}_i \boldsymbol{e}_j^T \right]^{\circ(\alpha-1)} \Big/ \|\boldsymbol{\gamma}\|_{1,\alpha}^{\alpha} \right)$;
  $\boldsymbol{S}^\star \leftarrow$ `Oracle I`$(\nabla_S J^k, \boldsymbol{S}^k)$; // Algorithm 2
  $\boldsymbol{\gamma}^\star \leftarrow$ `Oracle II`$(\nabla_{\boldsymbol{\gamma}} J^k, \boldsymbol{\gamma}^k)$; // Algorithm 3
  if $\langle \boldsymbol{S}^k - \boldsymbol{S}^\star, \nabla_S J^k \rangle_F + \langle \boldsymbol{\gamma}^k - \boldsymbol{\gamma}^\star, \nabla_{\boldsymbol{\gamma}} J^k \rangle_F \leq \epsilon$ **then**
    return $\boldsymbol{S}^k, \boldsymbol{\gamma}^k$;
  $\psi_k \leftarrow 2/(k+2)$;
  $\boldsymbol{S}^{k+1} \leftarrow (1-\psi_k)\boldsymbol{S}^k + \psi_k \boldsymbol{S}^\star$; // Frank–Wolfe updates
  $\boldsymbol{\gamma}^{k+1} \leftarrow (1-\psi_k)\boldsymbol{\gamma}^k + \psi_k \boldsymbol{\gamma}^\star$;
  $k \leftarrow k+1$;

In furtherance of attaining the first oracle (18), `Oracle I`, we note that it could be rewritten explicitly into the following *semidefinite programming (SDP)* problem on $\mathcal{S}^+$:

$$\begin{aligned} \text{Maximize} \quad & \operatorname{tr}\left[\nabla_S J(\boldsymbol{\gamma}, \boldsymbol{S})^T \widetilde{\boldsymbol{S}}\right] \\ \text{subject to} \quad & \operatorname{tr}\left[\boldsymbol{L}^+ \widetilde{\boldsymbol{S}}\right] \leq 1, \\ & \widetilde{\boldsymbol{S}} \succeq 0, \end{aligned} \tag{20}$$

in which $\boldsymbol{L}^+ \triangleq \sum_{\mathcal{L}^+} (\boldsymbol{x}_i - \boldsymbol{y}_j)(\boldsymbol{x}_i - \boldsymbol{y}_j)^T$. Due to its convexity, there exists several efficient algorithms for solving the optimum, including the most popular *interior point methods*. Yet, we take a different route towards the solution.

We find the core algorithm in [46] still applicable even if the objective of (20) varies from its origin in Lemma 1. It resorts to the simple gradient ascent method: $\widetilde{\boldsymbol{S}} \leftarrow \widetilde{\boldsymbol{S}} + \eta \nabla_{\widetilde{\boldsymbol{S}}} \operatorname{tr}[\nabla_S J(\boldsymbol{\gamma}, \boldsymbol{S})^T \widetilde{\boldsymbol{S}}]$, while in the meantime, the gradient of the objective function can be simplified using

$$\nabla_{\widetilde{\boldsymbol{S}}} \operatorname{tr}\left[\nabla_S J(\boldsymbol{\gamma}, \boldsymbol{S})^T \widetilde{\boldsymbol{S}}\right] = \nabla_{\widetilde{\boldsymbol{S}}} \left\langle \nabla_S J(\boldsymbol{\gamma}, \boldsymbol{S}), \widetilde{\boldsymbol{S}} \right\rangle_F = \nabla_S J(\boldsymbol{\gamma}, \boldsymbol{S}) = \sum_{\boldsymbol{x}_i \neq \boldsymbol{y}_j} \frac{\gamma_{ij} (\boldsymbol{x}_i - \boldsymbol{y}_j)(\boldsymbol{x}_i - \boldsymbol{y}_j)^T}{\sqrt{(\boldsymbol{x}_i - \boldsymbol{y}_j)^T \boldsymbol{S} (\boldsymbol{x}_i - \boldsymbol{y}_j)}}. \tag{21}$$

Every time when an update is received, extra projections are still required to ensure the feasibility. Set up with respect to the two constraints in (20), the projections are $\arg\min_{\boldsymbol{S}} \{\|\boldsymbol{S} - \widetilde{\boldsymbol{S}}\|_F \mid \operatorname{tr}[\boldsymbol{L}^+ \boldsymbol{S}] \leq 1\}$ and $\arg\min_{\boldsymbol{S}} \{\|\boldsymbol{S} - \widetilde{\boldsymbol{S}}\|_F \mid \boldsymbol{S} \succeq 0\}$, where the *Frobenius norm* $\|\cdot\|_F$ is drawn for gauging the distance between matrices. Merging these concepts of gradient ascent and iterative projection, `Oracle I` (Algorithm 2) is guaranteed an $O(m^2(+n^3))$ computational expense with a constant n. [3]

`Oracle II` (19) leading to $\boldsymbol{\gamma}^\star$ happens to be more sophisticated because of its underlying constrained set $\Gamma(\mu, \nu)$ which forms a convex polytope known as the *Birkhoff polytope* denoted as $\mathcal{B}$, whereas we would like to seek for an optimization oracle in the form of $\boldsymbol{\gamma}^\star = \arg\min_{\widetilde{\boldsymbol{\gamma}} \in \mathcal{B}} \operatorname{tr}(\boldsymbol{A}\widetilde{\boldsymbol{\gamma}})$. This could be achieved by our developed Algorithm 3 exploiting *Birkhoff–von Neumann theorem* [2]:

[3] Specifically, the first projection involves minimizing a quadratic objective over linear constraint in $O(n^2)$ while the second requires eigendecompostion in $O(n^3)$. Please resort to Xing *et al.* [46] for further details of the projections.

Algorithm 2: `Oracle I` (Gradient Ascent + Iterative Projection)

Data: $X = \{x_i\}_{i=1}^m$, $Y = \{y_j\}_{j=1}^m$, $\mathcal{L}^+$, $\nabla_S J$, $\gamma \in \mathcal{B}$, $S \succ 0$, $\epsilon, \eta > 0$
Result: $S^\star \succ 0$
$L^+ \leftarrow \sum_{\mathcal{L}^+} (x_i - y_j)(x_i - y_j)^T$;
$\nabla_{\widetilde{S}} \operatorname{tr}\left[\nabla_S J(\gamma, S)^T \widetilde{S}\right] \leftarrow \nabla_S J = \sum_{x_i \neq y_j} \gamma_{ij} (x_i - y_j)(x_i - y_j)^T \Big/ \sqrt{(x_i - y_j)^T S (x_i - y_j)}$;
while not converge **do**
  while $\widetilde{S}$ not converge **do**
    $\widetilde{S} \leftarrow \arg\min_S \left\{ \left\| S - \widetilde{S} \right\|_F \middle| \operatorname{tr}[L^+ S] \leq 1 \right\}$; `// Affine half-space projection`
    $\widetilde{S} \leftarrow \arg\min_S \left\{ \left\| S - \widetilde{S} \right\|_F \middle| S \succeq 0 \right\}$; `// Semidefinite cone projection`
  $\widetilde{S} \leftarrow \widetilde{S} + \eta \nabla_{\widetilde{S}} \operatorname{tr}\left[\nabla_S J(\gamma, S)^T \widetilde{S}\right]$; `// Gradient ascent`
$S^\star \leftarrow \widetilde{S}$;
return $S^\star$;

Lemma 4 (Birkhoff–von Neumann theorem). *The set of $m \times m$ doubly stochastic matrices forms a convex polytope in m^2-dimensional Euclidean space, with $m \times m$ permutation matrices as its vertices.*

Lemma 4 reflects that one can always decompose a doubly stochastic matrix $\widetilde{\gamma}$ into permutation matrices P_h weighted by constants $\theta_h \geq 0$ that sum to 1 (i.e. $\sum_h \theta_h = 1$). Let $\{P_h \mid \forall 1 \leq h \leq h_{\max}\}$ be the set of all $h_{\max}$ ($h_{\max} \leq m^2$) $m \times m$ permutation matrices and $a_h \triangleq \operatorname{tr}(AP_h)$ form a vector a,

$$\min_{\widetilde{\gamma} \in \mathcal{B}} \operatorname{tr}(A\widetilde{\gamma}) \approx \min_{\sum_h \theta_h = 1, \theta_h \geq 0} \operatorname{tr}\left[A\Big(\sum_h \theta_h P_h\Big)\right] = \min_{\sum_h \theta_h = 1, \theta_h \geq 0} \sum_h \operatorname{tr}(AP_h)\theta_h = \min_{\|\theta\|_1 = 1, \theta \geq 0} \langle a, \theta \rangle. \quad (22)$$

So, solving the optimum $\theta^\star$ of the linear program constrained by probability simplex Δ at the right-hand-side of (22) leads us to

$$\gamma^\star \approx \sum_h \theta_h^\star P_h. \quad (23)$$

On top of that, a recent advance [44] revisits Birkhoff's approach and provides an ϵ-approximate (i.e. arbitrarily close to optimum) algorithm `Birkhoff+` with at most $O(\log(1/\epsilon))$ permutation matrices, implying the existence of an $O(m \log(1/\epsilon))$ oracle by only a modest trade-off in optimality.

Pseudocode of `Birkhoff+` are shown in Algorithm 4, including the most important subroutines `SimplexLP`($\lfloor \gamma_h - \gamma \rceil^T, \mathcal{B}$) ($\lfloor \cdot \rceil$ is the nearest integer function) and `BirkhoffStep`(γ, γ_h, P_h). The former insists on solving another linear program using *simplex method*:

$$\begin{aligned} \text{Maximize} \quad & \operatorname{tr}\left[\lfloor \gamma_h - \gamma \rceil^T P_h\right] \\ \text{subject to} \quad & P_h \in \mathcal{B}, \end{aligned} \quad (24)$$

while the latter aims to find the minimum element of matrix $(\gamma - \gamma_h - J_m) \circ P_h$, in which J_m is the all-ones matrix of size $m \times m$. For further details, please resort to Valls, Iosifidis, and Tassiulas [44].

4.4.3 Computational Complexity. `SteepOTVR` (Algorithm 5) is a straightforward amendment of `Minimax Frank-Wolfe` for large-scale recommendations. *Since each recommended event should not be too distant from its origin, it is reasonable that we can cut down the scale of the optimization problem (as well as the size of coupling matrix γ) by restricting its underlying sample space.* In other words, we can address the problem without affecting the quality of the recommendation by solving

a series of separate subproblems on $\mathcal{X}_S \times \mathcal{Y}_R$, where $\coprod_S \mathcal{X}_S = \mathcal{X}$ and $\coprod_R \mathcal{Y}_R = \mathcal{Y}$, with $\coprod$ denoting the *disjoint union*.

Algorithm 3: `Oracle II` (Birkhoff–von Neumann theorem + `Birkhoff+`)

Data: $\nabla_{\boldsymbol{\gamma}} J$, $\boldsymbol{\gamma} \in \mathcal{B}$, $S \geq 0$, $h_{\max} = \lceil c \log(1/\epsilon) \rceil$, $c > 0$, $\epsilon > 0$
Result: $\boldsymbol{\gamma}^\star \in \mathcal{B}$
$\boldsymbol{A} \leftarrow \nabla_{\boldsymbol{\gamma}} J^T = \sum_{i,j} \gamma_{ij} \sqrt{(\boldsymbol{x}_i - \boldsymbol{y}_j)^T \boldsymbol{S} (\boldsymbol{x}_i - \boldsymbol{y}_j)} \boldsymbol{e}_j \boldsymbol{e}_i^T + \zeta \frac{\log_2 e}{1-\alpha} \left(\alpha \left[\sum_{i,j} \| \boldsymbol{e}_j^T \boldsymbol{\gamma} \|_1 \boldsymbol{e}_j \boldsymbol{e}_i^T \right]^{\circ(\alpha-1)} / \| \boldsymbol{\gamma} \|_{1,\alpha}^{\alpha} \right)$;
$(\boldsymbol{P}_1, \cdots, \boldsymbol{P}_{h_{\max}}) \leftarrow$ `Birkhoff+`$(\boldsymbol{\gamma}, \epsilon)$;
$h \leftarrow 1$;
$\boldsymbol{\gamma}^\star \leftarrow \boldsymbol{0}_{m \times m}$;
$\boldsymbol{a} \leftarrow [\mathrm{tr}(\boldsymbol{A}\boldsymbol{P}_1), \cdots, \mathrm{tr}(\boldsymbol{A}\boldsymbol{P}_{h_{\max}})]^T$;
$\boldsymbol{\theta}^\star \leftarrow$ `SimplexLP`$(\boldsymbol{a}^T, \Delta)$;
while $h \leq h_{\max}$ **do**
  $\boldsymbol{\gamma}^\star \leftarrow \boldsymbol{\gamma}^\star + \theta_h^\star \boldsymbol{P}_h$;
  $h \leftarrow h + 1$;
return $\boldsymbol{\gamma}^\star \in \mathcal{B}$;

Algorithm 4: `Birkhoff+` (ϵ-approximate Birkhoff decomposition, Valls *et al.* [44])

Data: $\boldsymbol{\gamma} \in \mathcal{B}$, $h_{\max} = \lceil c \log(1/\epsilon) \rceil$, $c > 0$, $\epsilon > 0$
Result: $(\boldsymbol{P}_1, \cdots, \boldsymbol{P}_{h_{\max}})$
$\boldsymbol{\gamma}_h \leftarrow \boldsymbol{0}_{m \times m}$;
$h \leftarrow 1$;
while $\| \boldsymbol{\gamma}_h - \boldsymbol{\gamma} \|_F > \epsilon$ and $h \leq h_{\max}$ **do**
  $\boldsymbol{P}_h \leftarrow$ `SimplexLP`$(\lfloor \boldsymbol{\gamma}_h - \boldsymbol{\gamma} \rceil^T, \mathcal{B})$;
  $\theta_h \leftarrow$ `BirkhoffStep`$(\boldsymbol{\gamma}, \boldsymbol{\gamma}_h, \boldsymbol{P}_h) + 1$;
  $\boldsymbol{\gamma}_{h+1} \leftarrow \boldsymbol{\gamma}_h + \theta_h \boldsymbol{P}_h$;
  $h \leftarrow h + 1$;
return $(\boldsymbol{P}_1, \cdots, \boldsymbol{P}_{h_{\max}})$;

Algorithm 5: `SteepOTVR` (scaling up for large dataset)

Data: $\boldsymbol{X} = \{\boldsymbol{x}_i\}_{i=1}^m$, $\boldsymbol{Y} = \{\boldsymbol{y}_j\}_{j=1}^m$, $\mathcal{X}$, $\mathcal{Y}$, $\boldsymbol{\gamma}^0 \in \mathcal{B}$, $\boldsymbol{S}^0 \geq 0$, $\alpha > 1$, $\zeta \gg \epsilon > 0$, $k_{\max} \in \mathbb{N}$
$\coprod_S \mathcal{X}_S \leftarrow \mathcal{X}$; `// We partition probability spaces`
$\coprod_R \mathcal{Y}_R \leftarrow \mathcal{Y}$; `// pursuant to desired granularity`
while S **do**
  while R **do**
    `Minimax Frank-Wolfe`$(\mathcal{X}_S, \mathcal{Y}_R)$;
    Calculate the cache size (by Rényi entropy) and (soft) cache hit ratio of each region;

Furthermore, we roll out the analyses of computational complexities of these algorithms:

- `Oracle I`: Calculation of the gradient in the objective consumes time resource quadratically, while the projection onto the positive semidefinite cone requires an eigendecomposition in cubic-time. These altogether lead to a complexity of $O(m^2(+n^3))$, in which $n \ll m$ is negligible.

- `Oracle II`: As for the second optimization oracle, the bottleneck lies in the gradient calculation of quadratic-time. The scaling factors $\mathrm{tr}(\boldsymbol{A}\boldsymbol{P}_k)$ of the objective function appears to be $O(m\log(1/\epsilon))$ in time, which greatly improves the original oracle of $O(m^3)$ (Table 1 in [16]).
- `Minimax Frank-Wolfe`: A combination of the two oracles induces a complexity of $O(m^2(+n^3))$.
- `SteepOTVR`: The complexity is $O(m^2/M)$ when $|S| = |R| = M$. See explanations in Appendix B.4.

Besides, we also give further metaheuristics in Appendix B.4 for pruning potential redundancies.

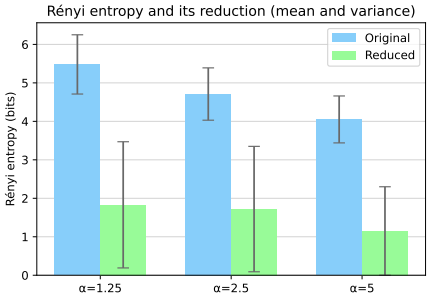
5 EVALUATION

Using the real world city-scale video viewing dataset described earlier in Section 2.2, we evaluate the performance of our proposed optimal transport video recommender and the `SteepOTVR` algorithm in this section. Further details on the evaluation are provided in Appendix B, including additional results such as the dynamics of `SteepOTVR` algorithm (Appendix B.5).

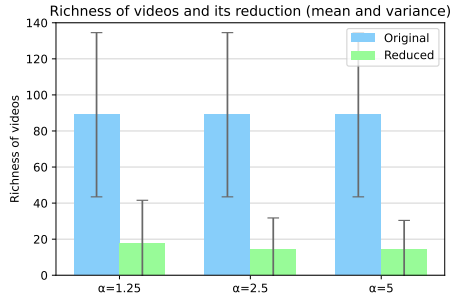
5.1 Reduction of Cache Sizes and VOD Traffic

The analytical result of Theorem 1 reveals a direct link between the soft cache hit ratio (SCHR) and Rényi entropy, where the latter is what we pick as a proxy of the cache sizes (Section 3.1). We justify the validity of choosing Rényi entropy as a proxy of cache sizes by experiments below:

- *Decreased Rényi Entropy.* Fig. 4a shows the Rényi entropy before and after the recommendation by 24 hours (a day), averaged over all regions connected to the same local cache and 31 days in December, 2014. In average, Rényi entropy reduced to around 31% of its original. The reduction of Rényi entropy corresponds to either the depletion of evenness or richness, with a trade-off controlled by parametrization of α. *Despite that evenness and richness may grow inversely, the long tail nature of video distribution circumvents us from this situation, when the plausible solution is to increase the overlaps by nudging the requests of niche videos to the popular ones.*



(a) The empirical Rényi entropies of the sets of videos watched over time, with α equals to 1.25, 2.5, and 5.



(b) The averaged size of the set of videos before and after recommendation. The blue bars are equivalent.

Fig. 4. Effects of video recommendation.

- *Decreased Cache Sizes.* `SteepOTVR` shrinks down the cardinality of the user-admissible set of videos, indicating both cache size and VOD traffic reduction. The green bars in Fig. 4b show the decrease of video richness (i.e. number of distinct videos) after recommendation, averaged hourly in December, 2014. Fig. 5 is an example of how the richness of videos varies over time. It is evident that sometimes the quantity of videos remain the same as the vertical lines highlight. This implies the recommendation performance may decay slightly during off-peak hours. We can plausibly infer that the proportion of niche content access increases during off-peaks, whereby finding a substitute with reduced cache size (and richness in videos) becomes more demanding.

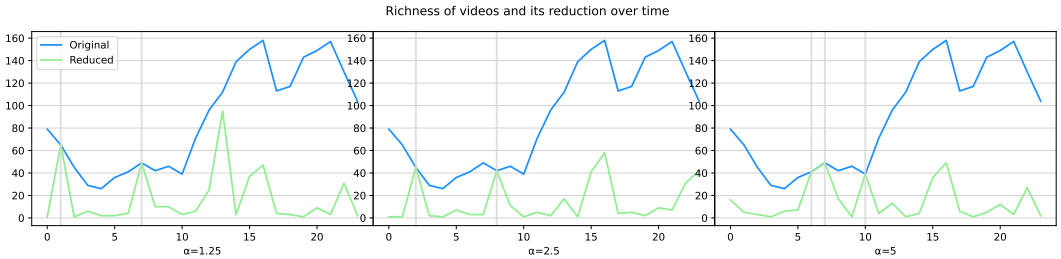


Fig. 5. Richness of videos over the timeline of Dec. 1, 2014 for different α, before and after recommendation.

5.2 Comparative Analysis

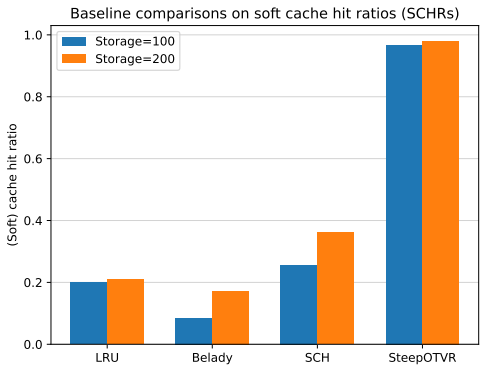
We pick three representative baselines, including the novel method of soft cache hits (SCH) [37] followed by two other classic cache replacement policies: the least recently used (LRU) [30] and the Bélády's optimal algorithm (Bélády) [1]. Detailed explanations of their respective mechanisms are rolled out in Appendix B.3.

- LRU. We compare first with the LRU, a widely used algorithm based on the heuristic that the recent accessed data are more likely to be accessed again in the near future.
- Bélády. The second baseline considered is the Bélády's optimal algorithm. Bélády operates by looking at the future memory access pattern of a given workload and evicts the cache content that will be accessed farthest in the future. This is known to be theoretically optimal [1], but as it relies on the knowledge of future access patterns, it is not possible to realise in practice (except for its various approximations including LRU, which can be seen as a heuristic to predict which item is not going to be needed in the near future). We use Bélády to compute the theoretical optimal performance achievable with a given cache size, assuming replacements through space-, time- or content-shifting (as proposed in this paper) are not allowed.
- SCH. As our third baseline, we use an adaptation of the original soft cache hit approach. There exist a variety of soft caching methods developed for various scenarios (e.g. [38, 39]). In general, they formulate cache placement problems on maximizing the SCHR (defined in respective forms), subject to binary constraints like $\sum_j b_j \leq B$. Each variable $b_j \in \{0, 1\}$ decides whether or not to cache the jth video, and B is the size of cache storage. *Whether or not a soft cache hit is counted depends on the predefined utility matrices of each individual user, storing the cost of shifting from one video to another. This is not only demanding but also less practical since it requires individual user preference data which are usually controlled by the content providers, not to mention the privacy concerns. On the contrary, our cache-aware recommendation method only needs to learn one distance metric $d_{S^\dagger}$ in terms of Mahalanobis distance.* Since individual preference is no longer required, our scheme aligns naturally to privacy-preserving designs.

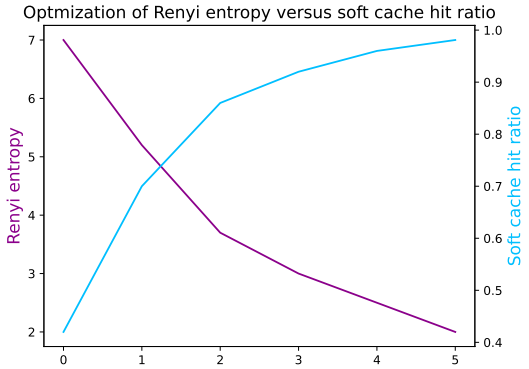
Fig. 6a shows that SteepOTVR significantly outperforms all the other baselines in terms of soft cache hits, and Fig. 6b facilitates empirical verification of the equivalence in Section 4.3, Theorem 1.

6 RELATED WORK

Starting with a broader horizon, information-centric network caching has pioneered the concept of unifying recommendations with caching [17, 31, 50]. An earlier but significant contribution in this field is presented in [40], where a learning-based approach was proposed for adaptively caching YouTube videos in a cellular network. This line of research has branched out into several directions, roughly classified as wireless Device-to-Device (D2D) caching and cooperative network caching. The former focuses on wireless scenarios (e.g. [22, 23, 32]), while the latter considers larger network scales involving infrastructures (e.g. [31, 41]). Both approaches highlight the potential



(a) Comparison of our SteepOTVR algorithm to the baseline through SCHRs.



(b) Experiment discloses that the SCHR is maximized when Rényi entropy is minimized.

Fig. 6. Evaluations on the soft cache hit ratios.

of edge caching [17, 19, 25, 29, 50]. Many studies in this area (e.g. [29, 44]) emphasize the concept of proximity clustering, leveraging contextual information about the features of the data. The exploitation of spatio-temporal information in caching consolidates where and when the users pull their requests [20, 48]. Noteworthy research by [6] aims to enlarge cache hits by nudging user preferences, paving a path towards *cache-aware recommendation*.

While content caching and recommendation may appear independent at the first glance, a shared intuition behind is to optimize a content-related objective while keeping the users satisfied to their recommendations. Under a tolerable distortion, [6] attempts to influence users' preferences for the purpose of increasing cache hits. Leveraging utility information of users, [37, 38] address cache placement problems maximizing the *soft cache hit ratio (SCHR)*. There also exist a number of extensions designed for specialized requirements or network schemes like real-time management [52], latency over non-orthogonal multiple access (NOMA) networks [13], and streaming experience [42, 53]. Another stream of work (e.g. [7, 49]) take a slightly different route, to conduct joint optimizations of caching and recommendation in small cell networks, either cooperatively or not. Yet, current literature (e.g. [13, 38, 52]) is not formulated in a data-driven nature, thereby overlooking potentially crucial featural characteristics such as spacetime statistics.

To the best of our knowledge, our paper develops the very first data-centric method that aims to shrink Internet VOD traffic by unifying caching and recommendation. This opens up the possibilities towards real implementations adaptive to the variation of users' tastes. *In comparison with the aforementioned papers, the major difference of ours is that we adapt our strategy from constructing deterministic discrete optimization problems (e.g. [6, 7, 33, 42]) to probabilistic models, so as to exploit the information which can be easily collected and updated after deployment.* The data-driven nature of our formulation implies its comfort in dealing with datasets at scale. It should also be capable of collaborating with various machine learning schemes as Appendix C.2 evinces.

7 CONCLUSION

We introduce a novel video recommender using optimal transport theory. SteepOTVR exploits overlapping spatio-temporal features of users' interests to significantly decrease the VOD caching traffic by optimizing Rényi entropy. This approach has the potential to enhance ICT-related environmental sustainability by shrinking traffic footprint, particularly in regards to the dominant video content delivery that decreases the CDN cache sizes and respective VOD traffic to the edge.

ACKNOWLEDGEMENTS

We thank Yanlin Chen (CWI Amsterdam) for sharing proving techniques, Caleb Wang (Northwestern University) and Leyang Xue (The University of Edinburgh) for environmental maintenance.

REFERENCES

[1] Laszlo A. Belady. 1966. A study of replacement algorithms for a virtual-storage computer. *IBM Systems Journal* 5, 2 (1966), 78–101.
[2] Garrett Birkhoff. 1946. Three observations on linear algebra. *Univ. Nac. Tacuman, Rev. Ser. A* 5 (1946), 147–151.
[3] Sem Borst, Varun Gupta, and Anwar Walid. 2010. Distributed Caching Algorithms for Content Distribution Networks. In *2010 Proceedings IEEE INFOCOM*. 1–9. https://doi.org/10.1109/INFCOM.2010.5461964
[4] Paul S Bradley, Kristin P Bennett, and Ayhan Demiriz. 2000. Constrained K-Means Clustering. *Microsoft Research, Redmond* (2000).
[5] Maarten A. Breddels and Jovan Veljanoski. 2018. Vaex: Big data exploration in the era of Gaia. *Astronomy & Astrophysics* (2018).
[6] Livia Elena Chatzieleftheriou, Merkouris Karaliopoulos, and Iordanis Koutsopoulos. 2017. Caching-Aware Recommendations: Nudging User Preferences Towards Better Caching Performance. In *IEEE INFOCOM 2017 - IEEE Conference on Computer Communications*. 1–9. https://doi.org/10.1109/INFOCOM.2017.8057031
[7] Livia Elena Chatzieleftheriou, Merkouris Karaliopoulos, and Iordanis Koutsopoulos. 2019. Jointly optimizing content caching and recommendations in small cell networks. *IEEE Transactions on Mobile Computing* 18, 1 (2019), 125–138. https://doi.org/10.1109/TMC.2018.2831690
[8] Cisco Systems, Inc. 2017. *CISCO VISUAL NETWORKING INDEX: FORECAST and TRENDS, 2017–2022*.
[9] Marco Cuturi. 2013. Sinkhorn Distances: Lightspeed Computation of Optimal Transport. In *Advances in Neural Information Processing Systems*, C.J. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K.Q. Weinberger (Eds.), Vol. 26. Curran Associates, Inc. https://proceedings.neurips.cc/paper/2013/file/af21d0c97db2e27e13572cbf59eb343d-Paper.pdf
[10] Sofien Dhouib, Ievgen Redko, Tanguy Kerdoncuff, Rémi Emonet, and Marc Sebban. 2020. A Swiss Army Knife for Minimax Optimal Transport. In *International Conference on Machine Learning*. PMLR, 2504–2513.
[11] U. Fano. 1947. Ionization yield of radiations. II. The fluctuations of the number of ions. *Physical Review* 72 (Jul 1947), 26–29. Issue 1. https://doi.org/10.1103/PhysRev.72.26
[12] Ronald A. Fisher. 1936. The use of multiple measurements in taxonomic problems. *Annals of Human Genetics* 7 (1936), 179–188. https://archive.ics.uci.edu/ml/datasets/iris
[13] Yaru Fu, Yue Zhang, Qi Zhu, Mingzhe Chen, and Tony Q. S. Quek. 2022. Joint content caching, recommendation, and transmission optimization for next generation multiple access networks. *IEEE Journal on Selected Areas in Communications* 40, 5 (2022), 1600–1614. https://doi.org/10.1109/JSAC.2022.3146901
[14] Aude Genevay. 2019. *Entropy-Regularized Optimal Transport for Machine Learning*. Ph. D. Dissertation. Paris Sciences et Lettres (ComUE).
[15] Aude Genevay, Gabriel Peyré, and Marco Cuturi. 2018. Learning Generative Models with Sinkhorn Divergences. In *International Conference on Artificial Intelligence and Statistics*. PMLR, 1608–1617.
[16] Martin Jaggi. 2013. Revisiting Frank-Wolfe: Projection-Free Sparse Convex Optimization. In *Proceedings of the 30th International Conference on Machine Learning (Proceedings of Machine Learning Research, Vol. 28)*, Sanjoy Dasgupta and David McAllester (Eds.). PMLR, Atlanta, Georgia, USA, 427–435. https://proceedings.mlr.press/v28/jaggi13.html
[17] Behrouz Jedari, Gopika Premsankar, Gazi Illahi, Mario Di Francesco, Abbas Mehrabi, and Antti Ylä-Jääski. 2021. Video caching, analytics, and delivery at the wireless edge: A survey and future directions. *IEEE Communications Surveys & Tutorials* 23, 1 (2021), 431–471. https://doi.org/10.1109/COMST.2020.3035427
[18] Lou Jost. 2006. Entropy and diversity. *Oikos* 113, 2 (2006), 363–375. http://www.jstor.org/stable/40234813
[19] Dmytro Karamshuk, Nishanth Sastry, Mustafa Al-Bassam, Andrew Secker, and Jigna Chandaria. 2016. Take-away TV: Recharging work commutes with predictive preloading of catch-up TV content. *IEEE Journal on Selected Areas in Communications* 34, 8 (2016), 2091–2101.
[20] Dmytro Karamshuk, Nishanth Sastry, Andrew Secker, and Jigna Chandaria. 2015. ISP-Friendly Peer-Assisted On-Demand Streaming of Long Duration Content in BBC iPlayer. In *2015 IEEE Conference on Computer Communications (INFOCOM)*. IEEE, 289–297.
[21] Dmytro Karamshuk, Nishanth Sastry, Andrew Secker, and Jigna Chandaria. 2015. On Factors Affecting the Usage and Adoption of a Nation-Wide TV Streaming Service. In *2015 IEEE Conference on Computer Communications (INFOCOM)*. 837–845. https://doi.org/10.1109/INFOCOM.2015.7218454
[22] Abdallah Khreishah, Jacob Chakareski, and Ammar Gharaibeh. 2016. Joint caching, routing, and channel assignment for collaborative small-cell cellular networks. *IEEE Journal on Selected Areas in Communications* 34, 8 (2016), 2275–2284. https://doi.org/10.1109/JSAC.2016.2577199

[23] Ming-Chun Lee, Mingyue Ji, Andreas F Molisch, and Nishanth Sastry. 2019. Throughput–outage analysis and evaluation of cache-aided D2D networks with measured popularity distributions. *IEEE Transactions on Wireless Communications* 18, 11 (2019), 5316–5332.

[24] Thomas Leinster and Mark W. Meckes. 2016. Maximizing diversity in biology and beyond. *Entropy* 18, 3 (9 March 2016). https://doi.org/10.3390/e18030088

[25] Dong Liu, Binqiang Chen, Chenyang Yang, and Andreas F. Molisch. 2016. Caching at the wireless edge: Design aspects, challenges, and future directions. *IEEE Communications Magazine* 54, 9 (2016), 22–28. https://doi.org/10.1109/MCOM.2016.7565183

[26] Jiaqiang Liu, Huan Yan, Yong Li, Dmytro Karamshuk, Nishanth Sastry, Di Wu, and Depeng Jin. 2021. Discovering and understanding geographical video viewing patterns in urban neighborhoods. *IEEE Transactions on Big Data* 7, 5 (2021), 873–884. https://doi.org/10.1109/TBDATA.2021.3055860

[27] Chi-Jen Roger Lo and Hung-Yun Hsieh. 2022. Information-Centric Scheduling: Queue Shortening in WSNs via Spatio-Temporal Compression. In *2022 Sensor Data Fusion: Trends, Solutions, Applications (SDF)*. 1–6. https://doi.org/10.1109/SDF55338.2022.9931951

[28] E Kathryn Morris, Tancredi Caruso, François Buscot, Markus Fischer, Christine Hancock, Tanja S Maier, Torsten Meiners, Caroline Müller, Elisabeth Obermaier, Daniel Prati, et al. 2014. Choosing and using diversity indices: Insights for ecological applications from the German Biodiversity Exploratories. *Ecology and Evolution* 4, 18 (2014), 3514–3524.

[29] Gianfranco Nencioni, Nishanth Sastry, Gareth Tyson, Vijay Badrinarayanan, Dmytro Karamshuk, Jigna Chandaria, and Jon Crowcroft. 2015. SCORE: Exploiting global broadcasts to create offline personal channels for on-demand access. *IEEE/ACM Transactions on Networking* 24, 4 (2015), 2429–2442.

[30] Elizabeth J. O'Neil, Patrick E. O'Neil, and Gerhard Weikum. 1993. The LRU-K Page Replacement Algorithm for Database Disk Buffering. In *Proceedings of the 1993 ACM SIGMOD International Conference on Management of Data* (Washington, D.C., USA) *(SIGMOD '93)*. Association for Computing Machinery, New York, NY, USA, 297–306. https://doi.org/10.1145/170035.170081

[31] Georgios S. Paschos, George Iosifidis, Meixia Tao, Don Towsley, and Giuseppe Caire. 2018. The role of caching in future communication systems and networks. *IEEE Journal on Selected Areas in Communications* 36, 6 (2018), 1111–1125. https://doi.org/10.1109/JSAC.2018.2844939

[32] Konstantinos Poularakis, George Iosifidis, Vasilis Sourlas, and Leandros Tassiulas. 2016. Exploiting caching and multicast for 5G wireless networks. *IEEE Transactions on Wireless Communications* 15, 4 (2016), 2995–3007.

[33] Kaiqiang Qi, Binqiang Chen, Chenyang Yang, and Shengqian Han. 2018. Optimizing Caching and Recommendation Towards User Satisfaction. In *2018 10th International Conference on Wireless Communications and Signal Processing (WCSP)*. 1–7. https://doi.org/10.1109/WCSP.2018.8555592

[34] Carl Edward Rasmussen and Christopher K. I. Williams. 2006. *Gaussian Processes for Machine Learning*. MIT Press. I–XVIII, 1–248 pages.

[35] Alfréd Rényi et al. 1961. On Measures of Entropy and Information. In *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability*, Vol. 1. Berkeley, California, USA.

[36] E. Schrödinger. 1932. Sur la théorie relativiste de l'électron et l'interprétation de la mécanique quantique. *Annales de l'institut Henri Poincaré* 2, 4 (1932), 269–310. http://eudml.org/doc/78968

[37] Pavlos Sermpezis, Theodoros Giannakas, Thrasyvoulos Spyropoulos, and Luigi Vigneri. 2018. Soft cache hits: Improving performance through recommendation and delivery of related content. *IEEE Journal on Selected Areas in Communications* 36, 6 (2018), 1300–1313. https://doi.org/10.1109/JSAC.2018.2844983

[38] Pavlos Sermpezis, Thrasyvoulos Spyropoulos, Luigi Vigneri, and Theodoros Giannakas. 2017. Femto-Caching with Soft Cache Hits: Improving Performance with Related Content Recommendation. In *GLOBECOM 2017 - 2017 IEEE Global Communications Conference*. 1–7. https://doi.org/10.1109/GLOCOM.2017.8254035

[39] Thrasyvoulos Spyropoulos and Pavlos Sermpezis. 2016. Soft Cache Hits and the Impact of Alternative Content Recommendations on Mobile Edge Caching. In *Proceedings of the Eleventh ACM Workshop on Challenged Networks* (New York City, New York) *(CHANTS '16)*. Association for Computing Machinery, New York, NY, USA, 51–56. https://doi.org/10.1145/2979683.2979688

[40] S. M. Shahrear Tanzil, William Hoiles, and Vikram Krishnamurthy. 2017. Adaptive scheme for caching YouTube content in a cellular network: Machine learning approach. *IEEE Access* 5 (2017), 5870–5881. https://doi.org/10.1109/ACCESS.2017.2678990

[41] Tuyen X. Tran and Dario Pompili. 2016. Octopus: A Cooperative Hierarchical Caching Strategy for Cloud Radio Access Networks. In *2016 IEEE 13th International Conference on Mobile Ad Hoc and Sensor Systems (MASS)*. 154–162. https://doi.org/10.1109/MASS.2016.029

[42] Dimitra Tsigkari and Thrasyvoulos Spyropoulos. 2022. An approximation algorithm for joint caching and recommendations in cache networks. *IEEE Transactions on Network and Service Management* 19, 2 (2022), 1826–1841. https://doi.org/10.1109/TNSM.2022.3150961

[43] Gareth Tyson, Nishanth Sastry, Richard Mortier, and Nick Feamster. 2016. Staggercast: Demand-Side Management for ISPs. *CoRR* abs/1605.09471 (2016). arXiv:1605.09471 http://arxiv.org/abs/1605.09471

[44] Víctor Valls, George Iosifidis, and Leandros Tassiulas. 2021. Birkhoff's decomposition revisited: Sparse scheduling for high-speed circuit switches. *IEEE/ACM Transactions on Networking* 29, 6 (Dec 2021), 2399–2412. https://doi.org/10.1109/TNET.2021.3088327

[45] C. Villani. 2008. *Optimal Transport: Old and New.* Springer Berlin Heidelberg. https://books.google.co.uk/books?id=hV8o5R7_5tkC

[46] Eric P. Xing, Andrew Y. Ng, Michael I. Jordan, and Stuart Russell. 2002. Distance Metric Learning, with Application to Clustering with Side-Information. In *Proceedings of the 15th International Conference on Neural Information Processing Systems (NIPS '02)*. MIT Press, Cambridge, MA, USA, 521–528.

[47] Kai Xu, Rajkarn Singh, Marco Fiore, Mahesh K. Marina, Hakan Bilen, Muhammad Usama, Howard Benn, and Cezary Ziemlicki. 2021. SpectraGAN: Spectrum Based Generation of City Scale Spatiotemporal Mobile Network Traffic Data. In *Proceedings of the 17th International Conference on Emerging Networking EXperiments and Technologies* (Virtual Event, Germany) *(CoNEXT '21)*. Association for Computing Machinery, New York, NY, USA, 243–258. https://doi.org/10.1145/3485983.3494844

[48] Taofeng Xue, Beihong Jin, Beibei Li, Weiqing Wang, Qi Zhang, and Sihua Tian. 2019. A Spatio-Temporal Recommender System for On-Demand Cinemas. In *Proceedings of the 28th ACM International Conference on Information and Knowledge Management* (Beijing, China) *(CIKM '19)*. Association for Computing Machinery, New York, NY, USA, 1553–1562. https://doi.org/10.1145/3357384.3357888

[49] Xiaolong Yang, Zesong Fei, Bin Li, Jianchao Zheng, and Jing Guo. 2022. Joint user association and edge caching in multi-antenna small-cell networks. *IEEE Transactions on Communications* 70, 6 (2022), 3774–3787. https://doi.org/10.1109/TCOMM.2022.3163763

[50] Jingjing Yao, Tao Han, and Nirwan Ansari. 2019. On mobile edge caching. *IEEE Communications Surveys & Tutorials* 21, 3 (2019), 2525–2553. https://doi.org/10.1109/COMST.2019.2908280

[51] Engin Zeydan, Ejder Bastug, Mehdi Bennis, Manhal Abdel Kader, Ilyas Alper Karatepe, Ahmet Salih Er, and Merouane Debbah. 2016. Big data caching for networking: Moving from cloud to edge. *IEEE Communications Magazine* 54, 9 (2016), 36–42. https://doi.org/10.1109/MCOM.2016.7565185

[52] Zhongyuan Zhao, Huihui Gao, Wei Hong, Xiaoyu Duan, and Mugen Peng. 2021. Joint design of content delivery and recommendation in wireless caching networks. *China Communications* 18, 11 (2021), 61–75. https://doi.org/10.23919/JCC.2021.11.005

[53] Dongsheng Zheng, Yingyang Chen, Mingxi Yin, and Bingli Jiao. 2020. Cooperative cache-aware recommendation system for multiple Internet content providers. *IEEE Wireless Communications Letters* 9, 12 (2020), 2112–2115. https://doi.org/10.1109/LWC.2020.3014266

APPENDIX

A More on Analyses

A.1 Dirac Delta. Data are always discrete due to the restriction of physical information storage, whereas the definiteness of probability densities only holds in absolute continuity. It can be useful in some situations to come up with a conceptual extension of the density. Dirac delta, as a generalised function (which is not a function), is formally written as:

$$\delta(x - x_0) = \begin{cases} 0, & x \neq x_0 \\ \infty, & x = x_0 \end{cases} \tag{25}$$

with a normalization where

$$\int_{\mathbb{R}} \delta(x)\mathrm{d}x = 1. \tag{26}$$

Further, through *Lebesgue's dominated convergence theorem*, for a sequence of Dirac deltas it holds:

$$\lim_{n\to\infty} \int_{-\infty}^{\infty} f(x)\delta_n(x - x_0)\mathrm{d}x = f(x_0). \tag{27}$$

Technically saying, the Dirac delta is neither a distribution nor a mathematical function, so the left-hand side of (27) should be seen as a limit but not a Riemann integral. Instead, it can be treated as a Stieltjes integral if desired, whereby $\delta(x)\mathrm{d}x$ integrates out as the Heaviside step function $\mathrm{H}(x)$.

A.2 *Proving Lemma 1.*

Proof. To start with, we first specify our discussion in the case where the set $\mathcal{L}^-$ contains all $m \times m$ data pairings $(\boldsymbol{x}_i, \boldsymbol{y}_j)$. Rewriting program (7) into an equivalent *epigraph* form, followed by the addition of dummy and slack variables $\beta_{ij} \geq 0$, $\delta_{ij} > 0$ that satisfy $\sum_{i,j} \beta_{ij} + \delta_{ij} = 1$, yields:

$$\begin{aligned} \text{Minimize} \quad & t \\ \text{subject to} \quad & \sum_{\mathcal{L}^+} (\boldsymbol{x}_i - \boldsymbol{y}_j)^T \Omega (\boldsymbol{x}_i - \boldsymbol{y}_j) \leq t, \\ & \sum_{i,j} \sqrt{(\boldsymbol{x}_i - \boldsymbol{y}_j)^T \Omega (\boldsymbol{x}_i - \boldsymbol{y}_j)} \geq 1, \\ & \sum_{i,j} \beta_{ij} + \delta_{ij} = 1, \ \beta_{ij} \geq 0, \ \delta_{ij} > 0, \ \ \forall 1 \leq i, \ j \leq m, \\ & \Omega \succeq 0. \end{aligned} \tag{28}$$

Observing that $\beta_{ij} + \delta_{ij} > \beta_{ij} \geq 0$ and $1 = \sum_{i,j} \beta_{ij} + \delta_{ij} > \sum_{i,j} \beta_{ij} \geq 0$ hold given the presence of the third constraint, it is permissible to introduce a new constraint

$$\sqrt{(\boldsymbol{x}_i - \boldsymbol{y}_j)^T \Omega (\boldsymbol{x}_i - \boldsymbol{y}_j)} \geq \beta_{ij} + \delta_{ij}, \ \ \forall 1 \leq i, \ j \leq m \tag{29}$$

as a valid replacement of the second constraint in (28). Additionally, by consolidating the third constraint in (28), we obtain $(\beta_{ij} + \delta_{ij})^{-1} > 0$ for each transportation pair.

Next, we transform the epigraph program using *Lagrange relaxation*, imposing multipliers $\xi_{ij} > 0$ on the new constraint (29) that is interchangeable with the second constraint of (28),

$$\begin{aligned} \text{Minimize} \quad & t + \sum_{i,j} \xi_{ij} \left[1 - (\beta_{ij} + \delta_{ij})^{-1} \sqrt{(\boldsymbol{x}_i - \boldsymbol{y}_j)^T \Omega (\boldsymbol{x}_i - \boldsymbol{y}_j)} \right] \\ \text{subject to} \quad & \sum_{\mathcal{L}^+} (\boldsymbol{x}_i - \boldsymbol{y}_j)^T \Omega (\boldsymbol{x}_i - \boldsymbol{y}_j) \leq t, \\ & \xi_{ij} > 0, \ (\beta_{ij} + \delta_{ij})^{-1} > 0, \ \ \forall 1 \leq i, \ j \leq m, \\ & \Omega \succeq 0. \end{aligned} \tag{30}$$

It is worthwhile to simplify $\xi_{ij} > 0$ and $(\beta_{ij} + \delta_{ij})^{-1} > 0$ by an alternative $\xi_{ij}(\beta_{ij} + \delta_{ij})^{-1} > 0$. Taking an opposite sign after dropping the constant part $t + \sum_{i,j} \xi_{ij}$ in the objective leads to

$$\begin{aligned} \text{Maximize} \quad & \sum_{i,j} \xi_{ij} (\beta_{ij} + \delta_{ij})^{-1} \sqrt{(\boldsymbol{x}_i - \boldsymbol{y}_j)^T \Omega (\boldsymbol{x}_i - \boldsymbol{y}_j)} \\ \text{subject to} \quad & \sum_{\mathcal{L}^+} (\boldsymbol{x}_i - \boldsymbol{y}_j)^T \Omega (\boldsymbol{x}_i - \boldsymbol{y}_j) \leq t, \\ & \xi_{ij} (\beta_{ij} + \delta_{ij})^{-1} > 0, \ \ \forall 1 \leq i, \ j \leq m, \\ & \Omega \succeq 0. \end{aligned} \tag{31}$$

Eventually, letting $S \triangleq \Omega / t \succeq 0$ and $\gamma_{ij} \propto \xi_{ij} (\beta_{ij} + \delta_{ij})^{-1} > 0$, with a normalization $\sum_{i,j} \gamma_{ij} = 1$ that assigns a probability measure, then we successfully disclose the equivalence of (7) and (8). □

A.3 *Proving Lemma 2.*

This appendix demonstrates the procedure of reformulating the Rényi entropy minimization problem in (5) into our core problem of Rényi-entropic optimal transport (9) in Lemma 2. These two programs are equivalent when the cost constraints in (5) entail Lemma 1.

Proof. Lagrange relaxation replaces the hard constraints $\mathcal{D}$ in (5) into the soft constraints in the objective. Selecting multipliers $\zeta^{-1}\gamma_{ij} \geq 0$ as the scaled probability of recommending $\boldsymbol{y}_j$ to $\boldsymbol{x}_i$, the Lagrange relaxation of problem (5) automatically excludes the case when $\gamma_{ij} = 0$, and writes:

$$\begin{aligned} \text{Minimize} \quad & \widehat{H}_\alpha(\boldsymbol{Y}) + \zeta^{-1} \sum_{i,j} \gamma_{ij} \left[\sqrt{(\boldsymbol{x}_i - \boldsymbol{y}_j)^T \boldsymbol{S}^\dagger (\boldsymbol{x}_i - \boldsymbol{y}_j)} - 1 \right] \\ \text{subject to} \quad & \boldsymbol{\gamma} \in \Gamma(\mu, \nu), \end{aligned} \tag{32}$$

with $\langle \cdot, \cdot \rangle_F$ regarding to *Frobenius inner product* and $\Gamma(\mu, \nu)$ being the set of all couplings $\boldsymbol{\gamma}$ having marginals μ and ν (on $\mathcal{X}$ and $\mathcal{Y}$, respectively). Despite that $\boldsymbol{\gamma} \in \Gamma(\mu, \nu)$ is not an explicit constraint in (5), it shall be legal to put it down here. One can simply think of $\boldsymbol{\gamma}$ as a change of variable from ν (or $\boldsymbol{q}$, the real variable) because μ (or $\boldsymbol{p}$) is assigned *a priori* when $\boldsymbol{X}, \mathcal{X}$ are given.

The distance metric $\boldsymbol{S}^\dagger$ parametrized by $\boldsymbol{S}$ is learned by enforcing an additional constraint $\boldsymbol{S} \in \mathcal{S}^+$ (Section 4.2), leading to the following minimax optimization problem:

$$\begin{aligned} \text{Minimize} \quad & \max_{\boldsymbol{S} \in \mathcal{S}^+} \langle \boldsymbol{C}, \boldsymbol{\gamma} \rangle_F + \zeta \widehat{H}_\alpha(\boldsymbol{Y}) \\ \text{subject to} \quad & \boldsymbol{\gamma} \in \Gamma(\mu, \nu). \end{aligned} \tag{33}$$

Obviously, the reformulated problem of optimal transport (9) has now been attained. It is easy to recognize the concavity of the inner maximization problem of (cost) metric learning.

Next, the entry-wise matrix $L_{p,q}$-norm of $\boldsymbol{A}$ with a_{ij} as its (i, j)-element is defined by:

$$\|\boldsymbol{A}\|_{p,q} \triangleq \left(\sum_j \left[\sum_i |a_{ij}|^p \right]^{q/p} \right)^{1/q}. \tag{34}$$

We observe by picking $p = 1$, $q = \alpha$, Rényi entropy (Definition 2) of $\boldsymbol{Y}$ can be rewritten as

$$\widehat{H}_\alpha(\boldsymbol{Y}) = \frac{1}{1-\alpha} \log_2 \left(\sum_j \left[\sum_i \gamma_{ij} \right]^\alpha \right) = \frac{1}{1-\alpha} \log_2 \left(\|\boldsymbol{\gamma}\|_{1,\alpha}^\alpha \right) = \frac{\alpha}{\alpha - 1} \left[-\log_2 \left(\|\boldsymbol{\gamma}\|_{1,\alpha} \right) \right]. \tag{35}$$

Notably, in the case of $\alpha > 1$, $\boldsymbol{\gamma} \geq 0$, both the entry-wise matrix norm $\|\cdot\|_{1,\alpha}$ and $-\log_2(\cdot)$ are convex. Consequently, the Rényi entropy $\widehat{H}_\alpha(\boldsymbol{\gamma})$ as their composition is convex with respect to $\boldsymbol{\gamma}$.

Lastly, from Lemma 4 we learn that the class of all $m \times m$ *doubly stochastic matrices* forms a *Birkhoff polytope*, while its convexity further indicates the convexity of the constrained set $\Gamma(\mu, \nu)$ since coupling $\boldsymbol{\gamma}$ could be any doubly stochastic matrix when the marginal ν is not specified. In addition, $\langle \boldsymbol{C}, \boldsymbol{\gamma} \rangle_F$ is affine on $\boldsymbol{\gamma}$. These altogether verify the convexity of (10) with regard to $\boldsymbol{\gamma}$. □

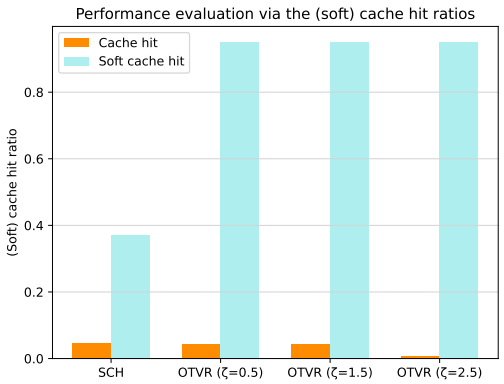


Fig. 7. Comparison of SCH and SteepOTVR under different regularizations in (soft) cache hit ratios.

At first glance, it might look like the parameter ζ in (9) introduces a trade-off between the optimal cache size (Rényi entropy) and the user satisfaction (cost metric). Nevertheless, the SCHR one can reach in theory should be 100%, because when emphasizing the Rényi entropy in (9) with a large ζ, the soft constraints in (32) would still remain satisfied (even for an infinitesimal $\zeta^{-1} > 0$). Specifying the values of ζ, Fig. 7 shows an empirical comparison of SCHRs between our approach and the SCH baseline, supporting the statement that ζ has only little influence on achievable SCHRs.

A.4 Proving Theorem 1.

Proof. We highlight the fact that $\log_2(\cdot)$ is a monotonically increasing function and the indicator function $\mathbb{1}[\cdot]$ is binary. The following inequality

$$\sum_j \Big[\sum_i \gamma_{ij} \mathbb{1}[c_{ij}^\dagger \le 1] \Big]^\alpha \le \sum_j \Big[\sum_i \gamma_{ij} \Big]^\alpha \tag{36}$$

leads to the first relation:

$$\widehat{H}_\alpha(\boldsymbol{Y}) \triangleq \frac{1}{1-\alpha} \log_2 \left(\sum_j \Big[\sum_i \gamma_{ij} \Big]^\alpha \right) \le \frac{1}{1-\alpha} \log_2 \left(\sum_j \Big[\sum_i \gamma_{ij} \mathbb{1}[c_{ij}^\dagger \le 1] \Big]^\alpha \right). \tag{37}$$

Simple derivations give us another relation below:

$$\sum_j \Big[\sum_i \gamma_{ij} \mathbb{1}[c_{ij}^\dagger \le 1] \Big]^\alpha \le \sum_j \Big[\sum_i \gamma_{ij} \mathbb{1}[c_{ij}^\dagger \le 1] \Big]^1 = \sum_{i,j} \gamma_{ij} \mathbb{1}[c_{ij}^\dagger \le 1] \triangleq \chi_{\text{SCH}}. \tag{38}$$

The first implies Rényi entropy minimization can be done equivalently through minimizing its upper bound, or maximizing the summation term in the logarithm. This coincides with the maximization of the left-hand side in (38) followed thereby the SCHR maximization at its right-hand side. □

A.5 Proving Lemma 3.

Proof. To start with, we note that the differentiability of $J(\boldsymbol{\gamma}, \boldsymbol{S})$ over $\boldsymbol{\gamma}$ and $\boldsymbol{S}$ can be verified by the classical epsilon-delta construction. So, the gradients at $(\boldsymbol{\gamma}, \boldsymbol{S})$ are

$$\begin{aligned} \nabla_{\boldsymbol{S}} J(\boldsymbol{\gamma}, \boldsymbol{S}) &= \frac{\partial J(\boldsymbol{\gamma}, \boldsymbol{S})}{\partial \boldsymbol{S}} = \frac{\partial \langle \boldsymbol{C}, \boldsymbol{\gamma} \rangle_F}{\partial \boldsymbol{S}} + \zeta \frac{\partial \widehat{H}_\alpha(\boldsymbol{\gamma})}{\partial \boldsymbol{S}}, \\ \nabla_{\boldsymbol{\gamma}} J(\boldsymbol{\gamma}, \boldsymbol{S}) &= \frac{\partial J(\boldsymbol{\gamma}, \boldsymbol{S})}{\partial \boldsymbol{\gamma}} = \frac{\partial \langle \boldsymbol{C}, \boldsymbol{\gamma} \rangle_F}{\partial \boldsymbol{\gamma}} + \zeta \frac{\partial \widehat{H}_\alpha(\boldsymbol{\gamma})}{\partial \boldsymbol{\gamma}}. \end{aligned} \tag{39}$$

Gradient $\nabla_{\boldsymbol{S}} J(\boldsymbol{\gamma}, \boldsymbol{S})$ is easier to derive since $\widehat{H}_\alpha(\boldsymbol{\gamma})$ is constant to $\boldsymbol{S}$ and there exists a simple form

$$\frac{\partial \langle \boldsymbol{C}, \boldsymbol{\gamma} \rangle_F}{\partial \boldsymbol{S}} = \sum_{i,j} \gamma_{ij} \frac{\partial \sqrt{(\boldsymbol{x}_i - \boldsymbol{y}_j)^T \boldsymbol{S} (\boldsymbol{x}_i - \boldsymbol{y}_j)}}{\partial \boldsymbol{S}} = \sum_{\boldsymbol{x}_i \neq \boldsymbol{y}_j} \frac{\gamma_{ij} (\boldsymbol{x}_i - \boldsymbol{y}_j)(\boldsymbol{x}_i - \boldsymbol{y}_j)^T}{\sqrt{(\boldsymbol{x}_i - \boldsymbol{y}_j)^T \boldsymbol{S} (\boldsymbol{x}_i - \boldsymbol{y}_j)}}. \tag{40}$$

On the other hand, the derivative of $\langle \boldsymbol{C}, \boldsymbol{\gamma} \rangle_F$ over $\boldsymbol{\gamma}$ is simply the cost matrix $\boldsymbol{C}$ owing to its affinity, and the succeeding term $\nabla_{\boldsymbol{\gamma}} \widehat{H}_\alpha(\boldsymbol{\gamma})$ can be re-expressed as follows:

$$\frac{\partial \widehat{H}_\alpha(\boldsymbol{\gamma})}{\partial \boldsymbol{\gamma}} = \frac{\log_2 e}{1-\alpha} \times \frac{\partial \log_e \big(\|\boldsymbol{\gamma}\|_{1,\alpha}^\alpha \big)}{\partial \boldsymbol{\gamma}} = \frac{\alpha \log_2 e}{1-\alpha} \times \frac{\big[\sum_{i,j} \|\boldsymbol{e}_j^T \boldsymbol{\gamma}\|_1 \boldsymbol{e}_i \boldsymbol{e}_j^T \big]^{\circ(\alpha-1)}}{\|\boldsymbol{\gamma}\|_{1,\alpha}^\alpha}, \tag{41}$$

with $\circ$ representing the Hadamard power (i.e. entry-wise power) emerged from the derivative

$$\frac{\partial \|\boldsymbol{\gamma}\|_{1,\alpha}^\alpha}{\partial \boldsymbol{\gamma}} = \frac{\partial \sum_j (\sum_i \gamma_{ij})^\alpha}{\partial \sum_i \gamma_{ij}} \times \frac{\partial \sum_i \gamma_{ij}}{\partial \gamma_{ij}} \times \frac{\partial \gamma_{ij}}{\partial \boldsymbol{\gamma}} = \alpha \left(\sum_i \gamma_{ij} \right)^{\alpha-1} \times \frac{\partial \gamma_{ij}}{\partial \boldsymbol{\gamma}} = \alpha \left[\sum_{i,j} \|\boldsymbol{e}_j^T \boldsymbol{\gamma}\|_1 \boldsymbol{e}_i \boldsymbol{e}_j^T \right]^{\circ(\alpha-1)}. \tag{42}$$

□

A.6 Smoothness and Compactness.

Lemma 5. *Our objective function J is L-smooth, namely, the gradient map $\nabla J \triangleq (\nabla_{\boldsymbol{\gamma}} J, \nabla_S J)$ is L-Lipschitz continuous in domain $\Gamma(\mu,\nu)\times\mathcal{S}^{++}$ where $\mathcal{S}^{++} \triangleq \{S \succeq \sigma \boldsymbol{I}\} \cap \mathcal{S}^{+}$ and $\sigma > 0$.*

Proof. Before delving into the proof, one should note that the substitution of $\boldsymbol{S}^+$ to $\boldsymbol{S}^{++}$ is a mild adjustment to circumvent undefined or unbounded gradient at the margin of the domain. This adaptation does not influence the solution (i.e. it is equivalent) and its convergence.

We start with manifesting that the cross-partial derivatives remain symmetric and continuous in the domain. The former can be verified by simple calculations

$$\nabla_{\boldsymbol{\gamma}}\nabla_S J(\boldsymbol{\gamma}, S) = \nabla_S \nabla_{\boldsymbol{\gamma}} J(\boldsymbol{\gamma}, S) = \sum_{\boldsymbol{x}_i \neq \boldsymbol{y}_j} \frac{(\boldsymbol{x}_i - \boldsymbol{y}_j)(\boldsymbol{x}_i - \boldsymbol{y}_j)^T}{\sqrt{(\boldsymbol{x}_i - \boldsymbol{y}_j)^T S (\boldsymbol{x}_i - \boldsymbol{y}_j)}}. \tag{43}$$

As for the latter, because the summation of continuous functions remains continuous, it is sufficient to show the continuity of each underlying term. By specifying

$$\delta = \epsilon \times \frac{2\sqrt{[(\boldsymbol{x}_i - \boldsymbol{y}_j)^T S (\boldsymbol{x}_i - \boldsymbol{y}_j)]^3}}{\|\boldsymbol{x}_i - \boldsymbol{y}_j\|_2^2 \sqrt{\sum_{i,j} \|\boldsymbol{x}_i - \boldsymbol{y}_j\|_2^4}}, \tag{44}$$

we can infer for any $\delta > 0$ and $0 < \|(S + \Delta S) - S\|_F < \delta$ there exists an $\epsilon > 0$ such that

$$\begin{aligned}
&\left\| \frac{(\boldsymbol{x}_i - \boldsymbol{y}_j)(\boldsymbol{x}_i - \boldsymbol{y}_j)^T}{\sqrt{(\boldsymbol{x}_i - \boldsymbol{y}_j)^T S (\boldsymbol{x}_i - \boldsymbol{y}_j)}} - \frac{(\boldsymbol{x}_i - \boldsymbol{y}_j)(\boldsymbol{x}_i - \boldsymbol{y}_j)^T}{\sqrt{(\boldsymbol{x}_i - \boldsymbol{y}_j)^T (S + \Delta S) (\boldsymbol{x}_i - \boldsymbol{y}_j)}} \right\|_F \\
&\qquad \leq \left\| \frac{(\boldsymbol{x}_i - \boldsymbol{y}_j)^T \Delta S (\boldsymbol{x}_i - \boldsymbol{y}_j)}{2\sqrt{[(\boldsymbol{x}_i - \boldsymbol{y}_j)^T S (\boldsymbol{x}_i - \boldsymbol{y}_j)]^3}} (\boldsymbol{x}_i - \boldsymbol{y}_j)(\boldsymbol{x}_i - \boldsymbol{y}_j)^T \right\|_F \\
&\qquad = \frac{(\boldsymbol{x}_i - \boldsymbol{y}_j)^T \Delta S (\boldsymbol{x}_i - \boldsymbol{y}_j)}{2\sqrt{[(\boldsymbol{x}_i - \boldsymbol{y}_j)^T S (\boldsymbol{x}_i - \boldsymbol{y}_j)]^3}} \times \sqrt{\sum_{i,j} \|\boldsymbol{x}_i - \boldsymbol{y}_j\|_2^4} < \epsilon.
\end{aligned} \tag{45}$$

Likewise, we can prove the continuity of the remaining second derivatives. Since continuity in a compact domain implies boundness, the L-smoothness of J comes up as an oblique consequence. □

Lemma 6. *The feasible set $\mathcal{D} \triangleq \{(\boldsymbol{\gamma}, S) | \Gamma(\mu,\nu)\times\mathcal{S}^{++}\}$ has a diameter of $D \leq 2\sqrt{m + \frac{\lambda_{\max}(S)}{\lambda_{\min}(\boldsymbol{L}^+)}}$.*

Proof. Let $\lambda_{\min}(\boldsymbol{A})$ as the smallest and $\lambda_{\max}(\boldsymbol{A})$ the largest eigenvalue of $\boldsymbol{A}$. Trace inequality $\lambda_{\min}(\boldsymbol{A})\mathrm{tr}(\boldsymbol{B}) \leq \mathrm{tr}(\boldsymbol{AB}) \leq \lambda_{\max}(\boldsymbol{A})\mathrm{tr}(\boldsymbol{B})$ valid for positive semidefinite matrices $\boldsymbol{A}, \boldsymbol{B}$ gives

$$1 \geq \mathrm{tr}(\boldsymbol{L}^+ S) \geq \lambda_{\min}(\boldsymbol{L}^+)\mathrm{tr}(S) > 0, \tag{46}$$

which further yields

$$\|S\|_F = \sqrt{\mathrm{tr}(S^T S)} \leq \sqrt{\lambda_{\max}(S^T)\mathrm{tr}(S)} \leq \sqrt{\frac{\lambda_{\max}(S^T)}{\lambda_{\min}(\boldsymbol{L}^+)}} = \sqrt{\frac{\lambda_{\max}(S)}{\lambda_{\min}(\boldsymbol{L}^+)}}. \tag{47}$$

Thus, a bound of D is right at our hands since for any $(\boldsymbol{\gamma} + \Delta\boldsymbol{\gamma}, S + \Delta S)$ and $(\boldsymbol{\gamma}, S)$ in $\mathcal{D}$, it holds that

$$\begin{aligned}
\|(\boldsymbol{\gamma} + \Delta\boldsymbol{\gamma}, S + \Delta S) - (\boldsymbol{\gamma}, S)\|_F^2 &= \|(\boldsymbol{\gamma} + \Delta\boldsymbol{\gamma}) - \boldsymbol{\gamma}\|_F^2 + \|(S + \Delta S) - S\|_F^2 \\
&\leq (\|\boldsymbol{\gamma} + \Delta\boldsymbol{\gamma}\|_F + \|\boldsymbol{\gamma}\|_F)^2 + (\|S + \Delta S\|_F + \|S\|_F)^2 \leq \left(2\sqrt{m}\right)^2 + \left(2\sqrt{\frac{\lambda_{\max}(S)}{\lambda_{\min}(\boldsymbol{L}^+)}}\right)^2.
\end{aligned} \tag{48}$$

□

B Further Evaluation Details and Results

B.1 Experimental Setup. All these analyses are implemented using a versatile high-level programming language Python with its library vaex. Vaex uses lazy out-of-core dataframes to explore large tabular datasets, executing basic statistical calculations on an n-dimensional grid up to a billion objects / rows per second [5]. However, since vaex utilizes lazy evaluation (i.e. call-by-need) that saves memory by avoiding redundant computations as well as a zero memory copy policy called memory mapping, another Python package pandas is also required for fetching hash keys. This usually becomes the bottleneck of computational speed.

B.2 Cache Allocation to Users. Following Section 2.3, Fig. 8 illustrates an example of how k-means clustering [4] allocates the users in different regions of Shanghai to their local caches.

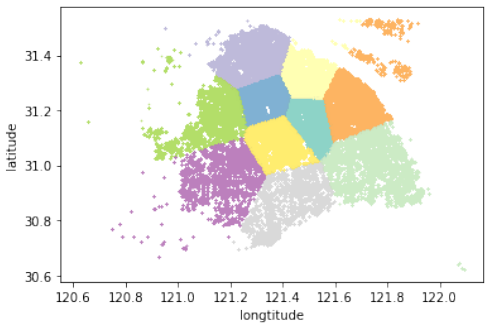


Fig. 8. Allocation of users to their local caches in terms of their video viewing events.

B.3 Details of the Baselines. This appendix presents details of SCH [37], LRU [30], and Bélády [1].

Starting from the foremost, the original series of work on soft cache hits (e.g. [38, 39]) has been integrated into a comprehensive paper: Sermpezis, Giannakas, Spyropoulos, and Vigneri [37], in which they formulate a cache placement problem to maximize the soft cache hit ratio (SCHR) under a binary constraint $\sum_j b_j \leq B$, with variables $b_j \in \{0, 1\}$ representing the decisions on whether or not to cache and B the cache size. The representative Problem 1 in [37] writes:

$$\begin{aligned} \text{Maximize} \quad & \sum_{u,i} p_i^u q_u \left[1 - \prod_j (1 - s_{ij}^u b_j)\right] \\ \text{subject to} \quad & \sum_j b_j \leq B, \\ & b_j \in \{0, 1\}, \end{aligned} \tag{49}$$

with the meanings of notations straightened out in Table 3. Problem 1 is, unfortunately, NP-hard. As a compromise, the authors present a greedy algorithm offering $(1 - 1/e)$-approximations.

p_i^u	probability of user u requesting content i
q_u	probability of user u in the range of the small cell (simply set to 1)
b_j	whether or not content j is stored in the local cache (of small cell)
s_{ij}^u	user u's utility when j is given as an alternative to the request of i

Table 3. Notations and their meanings. Opposed to [37], since user mobility is not of our interest, we simply set $q_u = 1$. Notice that the user-to-cache allocation is already well-defined in Section 2.2 and Appendix B.2.

- SCH. In SCH, the degree of acceptance is determined by utility matrices $\boldsymbol{S}^u$; whereas in our work, such a metric of acceptance is a learned cost. To set up the SCH baseline, we adapt the objective of (49) by replacing s_{ij}^u with $\Pr[c_{ij}^\dagger \leq 1]$. Unlike papers considering either mobility [39] or caching with helpers (i.e. femto-caching [38]), we omit the content request probabilities of individual users and stick to the viewpoint of local caches due to the essential difference of our objectives. Those settings are not vital at all since user-related probabilities they assume are just *ad hoc* parameters given in advance. Only the binary variables $b_j \in \{0, 1\}$ are crucial in regard to cache placement. Meanwhile, in order to give a fair comparison, information like the locations and timings of user requests are kept out of scope.
- LRU. The LRU algorithm ensures that the least recently used items are removed, which is often an effective strategy for optimizing cache performance. The core steps of LRU are as follows: Start with an empty cache and a recent list. When a request for data arrives, check if the data is in the cache. If it is, mark it as the most recently used item and return it. If the cache is full and a new item needs to be added, evict the item at the tail of the recency list (i.e. the least recently used) to make space for the new item. Whenever an item is accessed or a new item is added, move it to the head of the recency list to mark it as the most recently used.
- Bélády. The Bélády algorithm is considered optimal because it always selects the cache entry that minimizes the number of cache misses for a given workload, assuming complete knowledge of future access patterns. In practice, this requires knowledge of the entire future sequence of data accesses, which is not feasible. Below is the major procedure of Bélády: Initialize with an empty cache. As requests for data arrive, check if the data is already in the cache. If the answer is yes, then serve the request from the cache. Whenever the cache is full and a new request cannot be accommodated, Bélády predicts which item in the cache will be accessed farthest in the future (i.e. the cache entry that has the longest time until its next access). This entry is then evicted to make space for the new data. At last, the most critical step of the Bélády algorithm is the prediction of the future access patterns.

B.4 Hyperparameter Tuning. We lay out Table 4 as a crib for hypertunings. Sensitivity analyses in Section 5 and Appendix A.3 already reveal scenarios involving α and ζ, so our focus here is on the remaining parameters and potential considerations during adjustments.

Parameter	Purpose	Tuning metaheuristics
$\coprod_S \mathcal{X}_S$	Coarse-graining data	Provides a trade-off between speed and granularity ($\coprod_R \mathcal{Y}_R$)
α	Weighting diversity	Parametrises Rényi entropy regarding richness and evenness
ζ	Scaling multipliers	Increases as large as possible for emphasizing Rényi entropy
$\mathcal{L}^+(\boldsymbol{L}^+)$	Pairwise similarity	Sets featural similarity matrix $\boldsymbol{L}^+ \triangleq \sum_{\mathcal{L}^+}(\boldsymbol{x}_i - \boldsymbol{y}_j)(\boldsymbol{x}_i - \boldsymbol{y}_j)^T$

Table 4. A crib sheet for hyperparameter tuning.

Scaling up. We first explain how the computational complexity of SteepOTVR can be rewritten as $O(m^{3/2})$ by splitting the dataset into $M = \sqrt{m}$ subframes before feeding into the core algorithm Minimax Frank–Wolfe. According to the idea, we can break $\boldsymbol{X}$ down into M smaller subframes, each with size $\lceil m/M \rceil$, thereby the complexity becomes $O(M \times \lceil m/M \rceil^2) = O(m^2/M)$. In practice, by taking $M = \sqrt{m}$, the overall expense can be reduced to $O(m^{3/2})$. This is a plausible setting considering our CDN model and dataset are in city-scale. As an example, let's assume there are 100 caches in total serving their local access points over the city, and each event is allowed a 12-hour shifting period. Also, given that we have 141,598,427 rows in our dataset, this altogether splits the dataset into $M = 100 \times 60 \times 24/12 = 12000$ subframes — a quantity nearly equal to the square root of the total number of rows, $\sqrt{m} \approx \sqrt{141,598,427} \approx 11900$.

Distance metric learning. Although there are a variety of choices for the objective function and the constraints in (8), one should avoid choices that reduce the problem into a linear program, because solving such reduced problem can only lead to a rank 1 matrix $\boldsymbol{S}^\dagger$ which transfers no information. Therefore, the cost (distance metric) between videos $\boldsymbol{x}_i$ and $\boldsymbol{y}_j$ becomes $c^\dagger(\boldsymbol{x}_i, \boldsymbol{y}_j) \triangleq d_{\boldsymbol{S}^\dagger}(\boldsymbol{x}_i, \boldsymbol{y}_j)$ provided a learned optimal solution $\boldsymbol{S}^\dagger \succeq 0$. In addition, the '1' at the right-hand-side of the first constraint in (8) is sufficient without lost of generality. For any $t > 0$ alternative to 1, there is always a solution $t^2\boldsymbol{S}^\dagger \succeq 0$ that degenerates the constraint to its initial version [46].

For independent features with a diagonal $\boldsymbol{S}$, [46] suggests the use of *Newton–Raphson method*. They also suggests the solution for $\boldsymbol{S}$ being a full matrix, when the dependency among different columns can be captured in the distance we want to learn. Recalling Section 3.3.3, this aligns perfectly with our line of thought to jointly consider all features when doing recommendation. We modify (7), replacing $\mathcal{D}$ into the set of all possible pairs $(\boldsymbol{x}_i, \boldsymbol{y}_j)$. We further reset the notation of similar events as $\mathcal{L}^+$, which uniquely defines a linear similarity matrix $\boldsymbol{L}^+ \triangleq \sum_{\mathcal{L}^+}(\boldsymbol{x}_i - \boldsymbol{y}_j)(\boldsymbol{x}_i - \boldsymbol{y}_j)^T$. Such a similarity matrix plays an essential role in the trace constraint of (20):

$$\operatorname{tr}\left[\boldsymbol{L}^+\widetilde{\boldsymbol{S}}\right] \leq 1, \tag{50}$$

with $\widetilde{\boldsymbol{S}}$ being a dummy variable of $\boldsymbol{S}$. One can add in an extra trace inequality

$$\operatorname{tr}\left[\widetilde{\boldsymbol{S}}\right] \leq U \tag{51}$$

to enforce the compactness (i.e. closedness and boundedness) of constrained set. By raising the constant bound $U > 0$, the feasible region becomes closer to the original $\mathcal{S}^+$.

B.5 The Dynamics of Our Algorithm. Here we show the general applicability of our recommendation method through evaluations on the widely adopted classic: Fisher's Iris dataset [12].

Recall that a successful recommendation is to find some transportation plan which maps μ to ν, while ν is a steeper distribution that implies an increase on event overlaps captured by Rényi entropy. Such a transport map is achieved according to the iterative processes of our algorithm. To see how the distribution transforms over the iterations, we initialize $\nu^0 \triangleq \mu$ and let ν^k be the new distribution at the end of stage k in such way that $\nu^0 \leftrightarrow \nu^k \leftrightarrow \nu^{k+1} \leftrightarrow \nu$ form a Markov chain. All these distributions are defined purely on $\mathcal{Y} = \mathcal{X}$. By vectorizing these distributions we further write $\boldsymbol{\nu}^{k+1} = \boldsymbol{\gamma}^k\boldsymbol{\nu}^k$ for all $k \in \mathbb{Z}$ and $\boldsymbol{\nu}^k = \prod_{\ell=0}^{k-1}\boldsymbol{\gamma}^\ell\boldsymbol{\mu}$ for all $k \in \mathbb{N}$. Illustrated in Fig. 9, `Minimax Frank-Wolfe` gradually shifts the distribution from the original μ to a preferable optimal ν by updating the variables from $(\boldsymbol{\gamma}^k, \boldsymbol{S}^k)$ to $(\boldsymbol{\gamma}^{k+1}, \boldsymbol{S}^{k+1})$ in each iteration k.

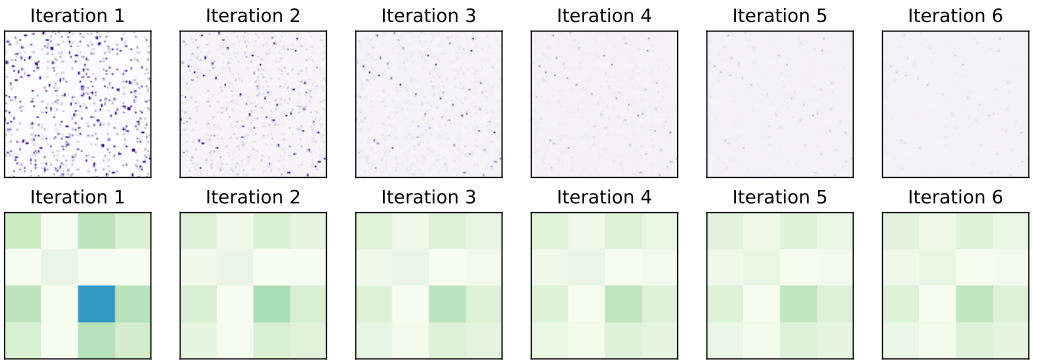
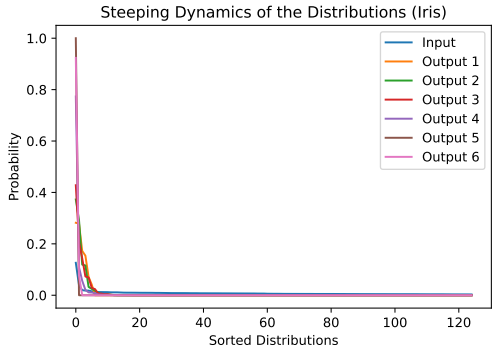
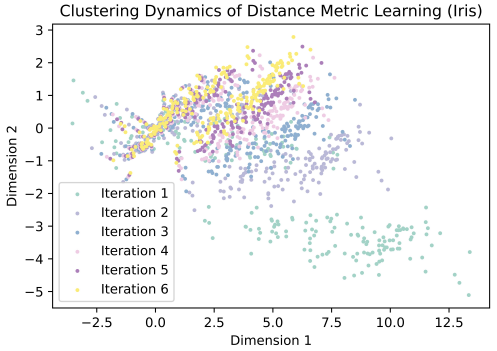


Fig. 9. The upper and lower layer shows $(\boldsymbol{\gamma}^k, \boldsymbol{S}^k)$ respectively using the widely adopted classic Fisher's Iris dataset [12]. However, it is hard to distinguish the (doubly stochastic) transport maps visually because of its fine-graining. Resulting effects are plotted in Fig. 10 and 11.

Fig. 10a and Fig. 10b plot respectively the dynamics of distribution steeping and distance metric learning with the Iris dataset, in which the data are rescaled by the square roots of $S^k \succeq 0$. Like Fig. 10a manifests, the distribution converges super fast, so we emphasize in Fig. 11 the variations of distributions over the first few iterations, by calculating the difference of distributions through simple subtractions. It can be seen that `SteepOTVR` converges close to the optimal after 6 iterations, with only a 5% difference at the peak (i.e. popular videos) of the distributions. Fig. 12 shows the clustering dynamics of distance metric learning on the Iris dataset through visually distinguishable projections from single 4-dimensional space onto its six 2-dimensional subspaces.



(a) To illustrate the steeping dynamics of the distribution, we start from a rather flat distribution, then make it steeper by solving (10), leading to distributions with smaller Rényi entropies.



(b) Distance metric learning and its clustering dynamics evaluated with 2-dimensional subspace projection. Similar data are forced to be closer to each other over iterations.

Fig. 10. The dynamics of distribution steeping and distance metric learning.

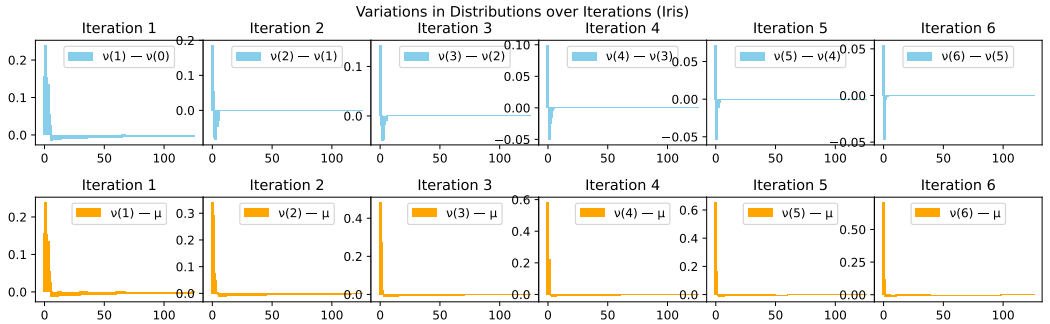


Fig. 11. The upper row depicts the difference of distributions by adjacent iterations, while the lower shows the variation of distribution compared to the initial. The distribution converges in just a few iterations.

B.6 Metaheuristics on Runtime Acceleration. `SteepOTVR` already offers an improvement on computational complexity, while the following data preprocessing metaheuristics might still come in handy for runtime acceleration, especially when dealing with large out-of-core datasets (e.g. 40 GB Shanghai dataset [26] applied in this paper):

- *Coarse-graining the sample space.* The abstraction of our statistical formulation is easily adaptable to different considerations of probability space we sample from. Its cardinality $|\mathcal{X}|$ implies the

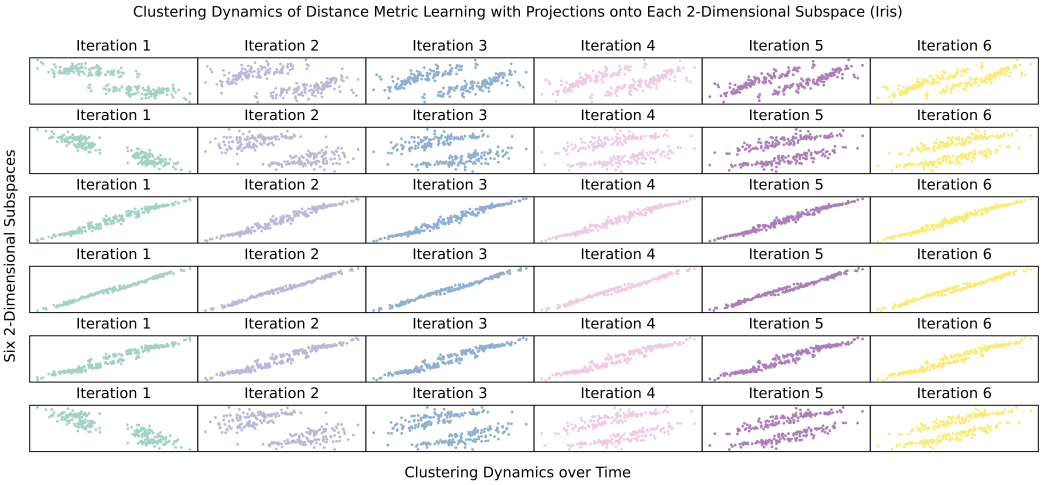


Fig. 12. Distance metric learning and its clustering dynamics evaluated on Fisher's Iris dataset, with projections onto all (six) 2-dimensional subspaces making it more visually distinguishable. It can be easily observed from these subspace projections that similar data are pushed closer to each other over iterations.

granularity of how we group and categorize the video viewing events, such as showing the distribution of video views by popularity. In parallel, before feeding the data into the recommender, we can first cluster similar events (e.g. closer in content and requested time) together in the granularity we want, and see them as a single event. For instance, providing timestamps precisely into seconds is not necessary at all in most real deployments. We can therefore exploit this property to reduce the input size, leading to the benefit of lower computational complexity.

- *Drop tail to emphasize the populars.* For events of unique or less popular videos (e.g. proactive searches or personalized, minor feeds), a plausible assumption of these user behaviors might be that they have already decided something specific to watch. Otherwise they would probably be watching those most popular videos on their feeds. Given that video recommendations are not coercive, even if we could successfully broadcast some sets of popular and satisfying videos to the edge, these users may still choose to stick to their determined searches or atypical tastes. This indicates that manipulating the videos out of favor is not effective to our objective at all. Coinciding with our previous statement of picking an $\alpha > 1$, it is the popular set of videos (or events) that should be emphasized. In correspondence, an even more straightforward mean for runtime amelioration is to simply neglect and discard the tails of distributions. To demonstrate why 'drop tail' can work, we set up an index called *k-contribution index*:

  Definition 5 (*k*-Contribution Index). *Denote the permutated vector of video views $w(\boldsymbol{X})$ as $\boldsymbol{w} \triangleq [w_1, ..., w_n]^T$ such that $w_i \geq w_{i+1}$ for each video i; and set $\boldsymbol{w}_{[k]}$ as the vector of the first k elements of $\boldsymbol{w}$. Then, we can define the k-contribution index as:*

$$\Delta\widehat{H}_\alpha^k \triangleq \frac{1}{1-\alpha}\log_2\left[\sum_{i=1}^{k}\left(\frac{w_i}{\|\boldsymbol{w}_{[k]}\|_1}\right)^\alpha \Big/ \sum_{i=1}^{k-1}\left(\frac{w_i}{\|\boldsymbol{w}_{[k-1]}\|_1}\right)^\alpha\right], \tag{52}$$

  where the vector 1-norm $\|\cdot\|_1$ implies the entry-wise summation of absolute values.

  It is evident from Fig. 13 and Definition 5 itself that the most popular videos, compared to the less, contribute more to the Rényi entropy which we aim to minimize, with a tendency controlled by α. This intuitively suggests the effectiveness of employing the 'drop tail' method as a metaheuristic

for reducing computational complexity. This approach is particularly useful in scenarios with limited resources, such as wireless sensor networks (WSNs) [27].

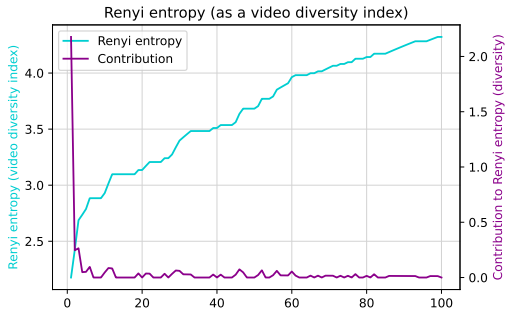


Fig. 13. Rényi entropy added up by the contribution indices, presented by percentage. Video with more views contributes more than the ones with less.

C Discussion and Future Work

C.1 Deployment Concerns. In order to cut down the VOD traffic through cache-aware recommendation, this paper gives a promising data-driven method, while we note that there are further aspects to be considered when it comes to real-world deployments:

- *Net neutrality.* Beyond designing operational mechanisms, we shall clarify that our recommender is not designed to 'control' user behaviors. It has to be admitted that techniques presented in this paper can pose concerns from a network neutrality perspective, as any recommendation system does. Yet, one may see it sanguinely as a pragmatic trade-off for traffic and cache size footprint shrinkage, and focus on the bright side and the positive influences (that the environmentally friendly nature of) our demand-side management could bring about.
- *User satisfaction.* The evaluation of user perception to recommendations is an important and foundational step before deployments. Nonetheless, partly due to the absence of personal information and our concentration on cache recommendation, user satisfaction and acceptance of recommendations have not been fully explored and addressed. It's worth noting that previous work *Staggercast* [43] did conduct a user study which broadly indicated a demand-side willingness for consuming alternate or time-shifted content, whereas we understand further studies are absolutely essential. Indeed, it is our plan for future work to do more in-depth user studies.
- *Business incentives.* Economical incentives is definitely a concern since it affects how the ecosystem evolves. Nowadays, recommendations are controlled by companies in application layer. VOD platforms like YouTube, Netflix, and social medias like Instagram, TikTok are now the mainstream, taking over a large sector of the market. Under current business models, nevertheless, only ISPs and telecom companies can benefit directly from the savings in operational cost due to decreased traffic. Better strategies are necessary to satisfy stakeholders in this business (e.g. ISPs, CDN service providers, VOD content providers, users, governing authorities, etc.).

C.2 Merging Cutting-Edge ML Models. Perhaps the most difficult challenge is the extension to involve predictions of future requests. While such an issue is not covered in this paper, the data-centric nature of our formulation endows a strong adaptability to miscellaneous state-of-the-art machine learning (ML) techniques:

- *Predictive data generation.* Generative adversarial networks (GANs) are unsupervised learning techniques in which a generator learns to discover patterns in the input data, enabling the model to generate new samples to augment the training data. The core idea of adversarial training

involves the structural mechanism of two neural networks competing against each other. These models improve themselves during the process with the aim to outperform their opponents. Several papers in machine learning have explored the application of optimal transport metrics as the loss function for generative models [14, 15]. As far as we recognize, existing methods address similarity learning and predictive video recommendation separately. It would be interesting to explore the potential of leveraging GANs to generate predictive data that apply to various network designs (e.g. [47]), especially those who align coherently with our means.

- *Learning in function space.* Another essential aspect of apart from learning and inference is to come up with precise predictions. Exploiting the technique of cross-validation across input data is one possible way to enhance prediction accuracy. On the other hand, *non-parametric models* may also bring some benefits through preserving higher fidelity and more information with functions. As an example, Mahalanobis distance can be easily extended into a non-linear regime by adding a mapping $\phi : \mathcal{X} \to \mathcal{H}$ from vector space to *Hilbert space*, written as:

$$d_S^{\phi}(\boldsymbol{x}, \boldsymbol{x}') \triangleq \sqrt{\phi(\boldsymbol{x} - \boldsymbol{x}')^T \boldsymbol{S} \phi(\boldsymbol{x} - \boldsymbol{x}')}, \quad \forall \boldsymbol{x}, \boldsymbol{x}' \in \mathcal{X}, \tag{53}$$

  where d_S^{ϕ} can be learned using modern statistical ML models like Gaussian processes (GPs) [34].

Received February 2023; revised January 2024; accepted January 2024