Individual Preference Probability Modeling for Video Content in Wireless Caching Networks

Ming-Chun Lee and Andreas F. Molisch
Ming Hsieh Department of Electrical Engineering
University of Southern California
Los Angeles, CA, USA
mingchul@usc.edu, molisch@usc.edu

Nishanth Sastry and Aravindh Raman
Department of Informatics
King's College London
London, UK
nishanth.sastry@kcl.ac.uk, aravindh.raman@kcl.ac.uk

***Abstract*—Caching of video files at the wireless edge, i.e., at the base stations or on user devices, is a key method for improving wireless video delivery. While *global* popularity distributions of video content have been investigated in the past, and used in a variety of caching algorithms, this paper investigates the *statistical modeling of the individual user preferences*. With individual preferences being represented by probabilities, we identify their critical features and parameters and propose a novel modeling framework as well as a parameterization of the framework based on an extensive real-world data set. Besides, an implementation recipe for generating practical individual preference probabilities is proposed. By comparing with the underlying real data, we show that the proposed models and generation approach can effectively characterize individual preferences of users for video content.**

I. Introduction

Data traffic generated by the demand for video content in wireless networks has approximately doubled every year and is expected to continue to grow in the next several years [1], [2]. Conventional approaches, such as using more efficient transceivers, densifying infrastructure, and/or using more spectrum, for supporting the increasing traffic are deemed insufficient or too expensive [2], [3]. An important alternative that has emerged in the past years is video caching at the wireless edge. Leveraging unique features of video popularity and the low cost of storage resources, video caching has shown its potential and drawn wide attention [1]–[4].

Video content caching has been discussed in different networks with different equipments being used as the storage resources [1]–[4]. Femtocaching and base station (BS) caching use storage resources in helper nodes and BSs to cache video content and provide the ability to immediately serve users without using backhaul [5]–[7]. Video content cached directly in mobile devices provides a more direct and a higher density caching apporach [8]–[10]. As the caching video content is in mobile devices, mobile users can either directly reach the video content from their own storage without consuming any resource [8] or exploit device-to-device communications to access video content with low cost [8]–[10]. The combination of storage on user devices together with coded multicast has also been widely explored [11].

Although designs for improving wireless video content caching have been widely explored, most of the literature adopts a homogeneous popularity model, i.e., assumes all users have the same file popularity distribution for deciding the desired video content [12]. Clearly this assumption violates the intuition that different users have different tastes and preferences. Therefore these designs are restricted to some extent due to lack of considering individual user preference. Note that modeling the individual preferences of a particular user, also known as the "Netflix challenge", has been investigated intensely [16], [17]. However, this is different from the need to find *statistics* of individual user distributions.

Approaches exploiting individual preference for caching or delivering content have just recently been discussed [12]–[16]. By exploiting individual preference, designs of wireless caching networks can be refined and improved [12]–[14]. Besides, analyses with individual preference can offer fundamental insights that might further enhance the system or strategy designs [15], [16]. However, to the best of our knowledge, there does not exist any statistical model for the individual user distributions based on real-world data. The current paper aims to fill this gap.

Our model uses hierarchies of probabilities to represent preferences of users. Empirically, video files can be categorized into genres according to their features, and users might have strong preferences toward a few genres [16]. The overall request probability of a user for a file is then modeled as the probability that a user wants a specific genre, and the popularity of a file within this genre. Since the individual preference probabilities of users can be described by the individual popularity distributions and ranking orders, statistics of them are respectvely investigated using the genre-based structure. We note that, in this paper, we implicitly denote the distribution as the rank-frequency distribution when we use the term: *popularity distribution*. We will extract the models and parameterization for these different statistics.

Such modeling and parameterization has to be based on real-world data to be meaningful. We are thus using data from an extensive dataset collected in the U.K. in 2014, namely the usage of the BBC iplayer [15], [16]. By observing the real data, we identify several important aspects of characterizing individual preferences, and propose the modeling framework for individual preference probabilities. By following the modeling framework, an individual preference probability generation approach is also proposed. We validate

the proposed modeling and generation approach with real data. The validation results demonstrate that proposed modeling and generation approach can effectively reproduce important features and statistcs of the individual preference. Therefore it can serve for designing, optimizing, analyzing, modeling, and simulating wireless caching networks.

The remaining paper is organized as follows. Section II introduces the basic modeling concepts and describes the necessary tools for manipulating the dataset. Main modeling results are provided in Sections III and IV. We propose the individual preference probability generation approach in Section V. Section VI presents the conclusions.

II. Individual Preference Probability Modeling and Dataset Preparations

A. Modeling on Individual Preference Probability

In this work, we consider the individual user probability distributions, which are defined as the probability that a specific user will in the future request a specific file for watching; multiple views by the same user are thus ignored (i.e., treated the same as single viewing). Since different users could have different preferences, preference probabilities of different users for the same file could be different. In this work we consider each file can be categorized into a genre, and there are G genres in the library. Therefore denoting M_g as the number of files in genre g, the total number of files in the library is given by $\sum_{g=1}^{G} M_g$. Given this library, we denote the preference probability of the file m in genre g for user k as $p_{g,m}^k$. Then the following properties must hold[1]: $0 \leq p_{g,m}^k \leq 1, \forall g, m, k$ and $\sum_{g=1}^{G} \sum_{m=1}^{M_g} p_{g,m}^k = 1, \forall k$.

To characterize individual preference probabilities of users, two important features need to be characterized: individual popularity distrbutions of files and individual ranking orders of files. Different individual popularity distrbutions represent different *concentration* rates of popularity distributions that different users might have, and different individual ranking orders represent different preferences for files by ranking files differently. Here we provide a simple example for elaboration. We consider two users with different preferences. Suppose that $G = 1$ and $M_1 = 3$. Therefore there are three files in the library. Then assume we know $p_{1,1}^1 = 0.5$, $p_{1,2}^1 = 0.3$, $p_{1,3}^1 = 0.2$; and $p_{1,1}^2 = 0.05$, $p_{1,2}^2 = 0.7$, $p_{1,3}^2 = 0.25$. Note that these six popularity values are a complete description, but obviously such a description becomes impossible to handle when considering thousands of files and millions of users. It can be observed that their popularity distributions are totally different. Besides, the ranking orders are different, namely 1, 2, 3 and 2, 3, 1, repsectively. By using the example, it can be observed that the differences between preferences of users can be fully described by the differences of individual popularity disributions and individual ranking orders.

[1]We note that, by using this model, we implicitly consider every user being equally important. However, from system's point of view, different weighting on different users according to certain strategy might be desired. Investigations of such weighted models from system's point of view are also important and are considered as a future direction.

To avoid confusions, in the following sections, we use global popularity/probability of genres/files to denote the popularity/probability of genres/files computed by taking all users into consideration. As the counterpart, the individual popularity/probability of genres/files is used to denote the popularity/probability computed by considering only a single specific user. In addition, without loss of generality, we consider the indices of genres to follow the decending order of the global popularities of genres, i.e., the global popularity of genre g is larger than the global popularity of genre $g+1$ for all $1 \leq g \leq G$.

B. Dataset Descriptions and Preprocessing

This work uses an extensive set of real-world data, namely the dataset of the BBC iplayer [15], [16]. The BBC iPlayer is a video streaming service provided by BBC (British Broadcasting Corporation). Video and radio content are provided for a number of BBC channels without charge. Content on the iPlayer is basically available for up to 30 days depending on the policies. We consider the month-long dataset accommodating up to 192,120,311 recorded access sessions of June, 2014. In each record, access information for the video content contains two important column: *user id* and *content id*. *user id* is based on the long-term cookies that unqiuely (in an anonymized way) identify users. *content id* is the specific identity that uniquely identifies each video content separately. Although there are certain exceptions, *user id* and *content id* can generally help us identify the user and the video content of each access. In addition to access identifications, video files in the BBC iplayer are annotated with one or more genres. The annotated genres for each video content are used to help us identify users' preferences. Notice that there are certain files that are not annotated with any genre. We simply filter them out, as described in the following paragraph. Detailed descriptions of the BBC iplayer dataset can be found in [15], [16].

To facilitate the investigation, preprocessing is conducted on the dataset. To be specific, we concentrate our investigation on regular (frequent) users. To define a regular user, we first define the unique access. By observations, we notice that a user could access the same file multiple times, possibly due to temporary disconnnections from Internet and/or due to temporary pauses raised by users when moving between locations. Since a user is generally unlikely to access the same video after finishing to watch the video within the period of a month, we consider multiple accesses made by the same user to the same file as a single unique access. A regular user is a user with more than 100 unique accesses in a month. An analysis that includes less frequent users will be presented in [19].

As described previously, a file could be annotated with one or several genres. The genre-wise classification is the foundation for characterizing preferences of users in our work. Hence if a file could not be classified into any genre, i.e., if no genre is annotatated on the file, the file is filtered out during the preprocessing.

C. Kullback-Leibler distance based parameter estimation

In Secs. III and IV, we propose models to fit statistics acquired from the dataset. To find the parameters that best fit the proposed models to the real data, the minimum Kullback-Leibler (K-L) distance approach is adopted and is given by

$$\hat{\mathbf{x}} = \arg\min_{\mathbf{x}} D_{KL}(\mathbf{x}) = \sum_{m} p_m^{real} \log \frac{p_m^{real}}{p_m^{model}(\mathbf{x})}, \quad (1)$$

where $\mathbf{x}$ is the vector representation of parameters, p_m^{real} is the probability of outcome m in real data, and $p_m^{model}(\mathbf{x})$ is the probability of outcome m characterized by the proposed model and $\mathbf{x}$. We note that $p_m^{real} \log \frac{p_m^{real}}{p_m^{model}(\mathbf{x})} = 0$ if $p_m^{real} = 0$ by definition; $\sum_m p_m^{real} = 1$; and $\sum_m p_m^{model}(\mathbf{x}) = 1$. We note that, to find a good fit of the target statistics, the following steps are basically used: (I) we choose distributions based on visual inspection; (II) we confirm the fitness of the chosen distributions by the above K-L test.

D. Genre-Based Structure and Modeling

In this work, a genre-based structure is adopted for the proposed modeling. This structure is adopted both for pragmatic and fundamental reasons. From a practical point of view, a direct modeling of individual popularities would involve too many parameters (a similar reasoning underlies, e.g., cluster-based modeling of wireless propagation channels). More fundamentally, it is infeasible to formulate the statistics of individual user preferences on files by simply observing the accesses of users[2]: in other words, a user does not have a *probability* to access a specific file - it either requests it or does not. Therefore, instead of directly finding the statistics of file preferences, we consider firstly investigating the statistics of genre preferences of users, and then approximating the file preferences within each genre by using the conditional non-user-specific statistics of files in each genre.

Since the perference probabilities of a user are fully described by its corresponding individual popularity disributions and ranking orders, we investigate statistics of the individual popularity distribution and ranking order using the genre-based structure in Secs. III and IV, respectvely. To provide a clear overview of the proposed modeling, a simple two-part summary is provided as follows.

Firstly, to characterize the statistics of the individual popularity distribution, we use the following distributions and models:

- Size distribution (Sec. III.A): since each user is only interested in a small number of genres, we use size distribution to indicate the statistics of *how many genres a user is watching*.
- Individual genre popularity distribution (Sec. III.B): given the number of desired genres for a user, individual genre popularity distribution characterizes *in which preference a user is watching a genre*.
- Genre-based conditional popularity distribution (Sec. III.C): we use the genre-based conditional popularity distribution of each genre to approximate the file popularity distribution *within the corresponding genre*.

Secondly, to characterize the statistics of the individual ranking order, we use the following distributions and models:

- Size distribution (Sec. III.A): the size distribution is again used here because it indicates *how many genres we need to rank for a user*.
- Genre appearance probabilities (Sec. IV.A): Since only the desired genres of a user need to be ranked, for a given genre, we use genre appearance probabilities to characterize *the possibilities of genres that are desired by a user*.
- Genre ranking distribution (Sec. IV.B): For a genre, its genre ranking distribution characterizes *the probability distribution in terms of rank for the genre* given that the genre is desired by the user.
- We directly use the global ranking order for files *within each genre* to approximate the individual ranking order for files within the corresponding genre.

To generate individual preference probabilities of a user from the proposed modeling, individual popularity distribution and ranking order are first generated by their corresponding distributions, respectively. Then by linking their results, the desired probabilities are generated. The proposed generation approach is elaborated and validated in Sec. V.

III. Proposed Modeling of Individual Popularity Distrbutions

Here the genre-based structure is adopted. The relevant statistics of genre popularity of users are firstly investigated. Then the genre-based conditional popularity distribution for files in each genre are investigated.

A. Size Distribution

Here the size distribution is investigated and modeled. By observations from real data, we found that a user would usually access a small number of genres even if there are more than one hundred genres in the library, and even though we consider users that access the iPlayer more than 100 times per month. These observations can be intuitively explained by that people usually have their own interests which constitute only a small portion of the whole world.

To quantify these observations, the size distribution[3] is investigated and modeled as

$$Pr(S_k = i) = \frac{f_i}{\sum_{j=1}^{M_s} f_j} \sim DouZipf(s_c, \gamma_1, \gamma_2, M_s) \quad (2)$$

[2]Certain user grouping approach might be feasible. However, the challenge is then becoming the designs of the group size and grouping approach.

[3]We model the number of genres accessed by the user as a random variable described by the size distribution. Therefore the size distribution is not user-dependent while different users could have different numbers of accessed genres.

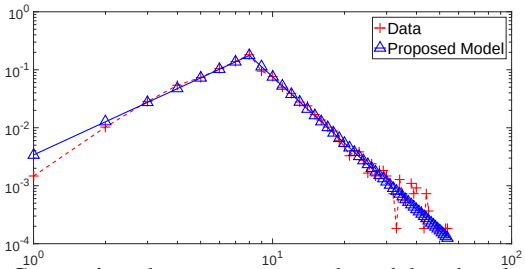


Fig. 1: Comparison between proposed model and real data for size distribution.

where $1 \leq i \leq M_s$; S_k is the number of genres being accessed by user k; s_c, γ_1, γ_2 are parameters that characterize the proposed modeling distribution; and

$$f_i = \begin{cases} \frac{i^{\gamma_1}}{s_c^{\gamma_1}}, \text{ if } i \leq s_c \\ \frac{i^{-\gamma_2}}{s_c^{-\gamma_2}}, \text{ if } i > s_c \end{cases}. \quad (3)$$

s_c specifically characterizes the point that offers the peak value of the modeling distribution; γ_1 and γ_2 characterize the ascending and descending behaviors of the distribution, respectively; M_s characterizes the maximal value of the random variable entailed by the distribution. We note that the proposed modeling distribution is named double-sided Zipf (DS-Zipf) distribution, because it behaves identical to a Zipf distribution if tracing from s_c to both 1 and M_s (it is, essentially, a discretized version of a double-sided exponential). We compare the proposed DS-Zipf modeling with the real distribution generated from the dataset in Fig. 1. Parameters in the figure are $s_c = 8, \gamma_1 = 1.9, \gamma_2 = 3.8, M_s = 54$, implying that each user watches typically only videos from 8 different genres. It can be observed that the proposed model is able to effectively reproduce the distribution generated by real data.

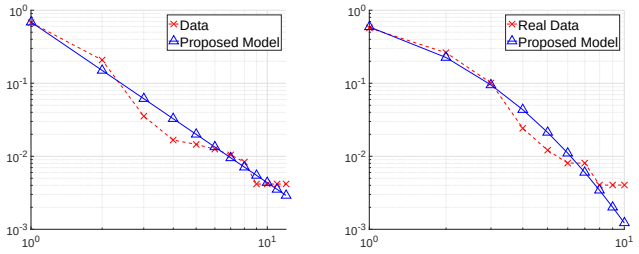
B. Individual Genre Popularity Distribution

The popularity of a genre g for a specific user k is defined as the ratio between the number of accesses to genre g by user k and the total number of accesses by the same user[4]. Therefore characterizing the individual genre popularity distribution is to characterize the difference between concentration levels of popularities in terms of genres. The proposed model for the individual genre popularity distribution is the Mandelbrot-Zipf (M-Zipf) distribution [18], expressed as

$$P_k^{out}(i) = \frac{\frac{1}{(i+q_k^{out})^{\gamma_k^{out}}}}{\frac{1}{\sum_{j=1}^{S_k}(j+q_k^{out})^{\gamma_k^{out}}}}, \quad (4)$$

where S_k is the number of genres accessed by user k, $P_k^{out}(i)$ is the popularity of the ith ranked genre, γ_k^{out} is the Zipf factor, and q_k^{out} is the plateau factor. In (4), parameters γ_k^{out} and q_k^{out} should follow a certain joint distribution. In the analyzed dataset, the range of γ_k^{out} is generally in $(0, 20]$; the range of q_k^{out} is generally in $(-1, 30]$. Notice that a specific user k would have a specific combination of γ_k^{out} and q_k^{out}. Therefore to fully describe γ_k^{out} and q_k^{out}, a statistical modeling for them is needed and is considered in [19].

[4]As some files are annotated with multiple genres, we consider each annotated genre accessed $\frac{1}{N}$ times when a file with N annotated genres is accessed.



(a) Case 1. (b) Case 2.

Fig. 2: Exemplary comparisons between proposed model and real data of individual genre popularity distributions.

In Fig. 2, we provide exemplary comparisons of the proposed model with distributions of real data. Parameters of the M-Zipf distribution are given as $\gamma_k^{out} = 2.2, q_k^{out} = 0$ and $\gamma_k^{out} = 7, q_k^{out} = 8.2$ for Figs. 2a and 2b, respectively. From both figures, it can be observed that the proposed M-Zipf model can effectively characterize real distributions. Note that this description does not specify *which* genre is the most popular one for this particular user; this aspect of genre ranking will be discussed in Sec. IV.B.

C. Genre-Based Conditional Popularity Distribution

The genre-based conditional popularity distribution of a given genre is the conditional probability distribution under the condition that files are annotated with the given genre. We use this distribution to approximate the *per-user* conditional preference probabilites of files under the condition that the file is annotated with the desired genre. We emphasize that the approximation is due to the impossibility of the direct characterization of user-based file preference statistics as discussed at the beginning of Sec. II.D. Since genre-based conditional popularity distributions are non-user-specific distributions, different users are assumed to have the same distribution within the same genre.

To model the genre-based conditional popularity distribution of genre g, we propose to again use M-Zipf distribution,

$$P_g^{in}(i) = \frac{\frac{1}{(i+q_g^{in})^{\gamma_g^{in}}}}{\frac{1}{\sum_{j=1}^{M_g}(j+q_g^{in})^{\gamma_g^{in}}}}, \quad (5)$$

where $P_g^{in}(i)$ is the popularity of the ith ranked file in genre g, γ_g^{in} is the Zipf factor, and q_g^{in} is the plateau factor. In the adopted dataset, the range of γ_g^{in} is generally in $(0, 20]$; the range of q_g^{in} is generally in $(-1, 1000]$. Due to the similar reason in Sec. III.B, statistical modeling for parameters in (5) is required and considered as future work [19]. In Fig. 3, the proposed model is compared with the real distribution of genre "factual". Parameters of the M-Zipf distribution is given as $\gamma_g^{in} = 2.4, q_g^{in} = 69$. From the figure, we observe that the proposed M-Zipf distribution can effectively model the real distribution.

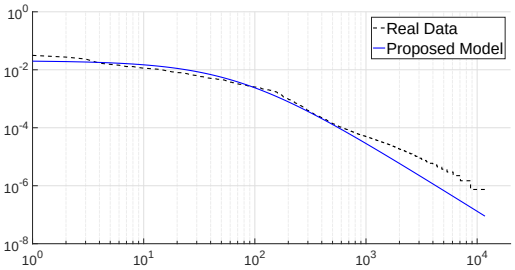


Fig. 3: Exemplary comparison between proposed model and real data of genre-based conditional popularity distribution.

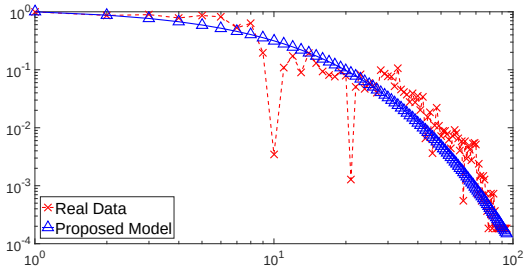


Fig. 4: Comparisons between proposed model and real data of genre apperance probabilities.

IV. Proposed Modeling of Individual Ranking Orders

In this section, statistics relevant to individual ranking order are investigated. Identical to the approach for investigating the individual popularity, a genre-based structure is adopted.

A. Genre Appearance Probability

As elaborated in previous sections, the number of genres that a user might access is usually much smaller than the total number of genres in the library. Therefore for each user k, we could obtain a genre list which is defined as the collection of all genres that are accessed by user k. The number of genres in the genre list of user k is by definition given by S_k.

Since the genre list of a user explicitly indicates the specific preference of that user for genres, characterizing statistics of the genre list is necessary. To characterize the corresponding statistics, the genre appearance probability is used. We define the appearance probability of genre g as the probability of genre g to appear in genre lists of users, and it is given by the ratio between the number of times that genre g appears in genre lists of users and the number of users. The proposed model describing genre appearance probabilities is[5]

$$P_{ap}(g) = \frac{(1+q_{ap})^{\gamma_{ap}}}{(g+q_{ap})^{\gamma_{ap}}}, \tag{6}$$

where $P_{ap}(g)$ is the appearance probability of genre g. Parameters γ_{ap} and q_{ap} are specifically given according to the dataset. In the adopted dataset, we have $\gamma_{ap} = 13.5$ and $q_{ap} = 100$. The comparison between the proposed model and the real data is provided in Fig. 4.

B. Genre Ranking Distribution

Given the genre list of a user, the ranking order of genres in the list also characterizes the preference of the user. To investigate the statistics of the ranking order, we investigate the ranking distributions of genres. The ranking distribution of a genre g is defined as the distribution of rank of genre g in genre lists of users conditioning on genre g appearing in those genre lists. By this definition, we denote $Pr(R_g = i)$ as the probability of genre g to be the ith ranked genre when genre g appears in a genre list. The proposed model for the distribution of this quantity is a DS-Zipf distribution,

$$Pr(R_g = i) = \frac{f^{rk}_{i,g}}{\sum_{j=1}^{G} f^{rk}_{j,g}} \sim DouZipf(s^{rk}_{c,g}, \gamma^{rk}_{1,g}, \gamma^{rk}_{2,g}, G). \tag{7}$$

The DS-Zipf distribution in (7) follows the same definition in (2) and (3). In the adopted dataset, the range of $s^{rk}_{c,g}$ is generally in $[1, 40]$; the range of $\gamma^{rk}_{1,g}$ is generally in $(0, 20]$; the range of $\gamma^{rk}_{2,g}$ is generally in $(0, 20]$. By empirical observations, we find that $s^{rk}_{c,g}$ generally increases as the rank of global popularity of genre g decreases, i.e., as g increases.

In Fig. 5, exemplary comparisons between the proposed model and the real data are provided. Parameters for Fig. 5a are $s^{rk}_{c,g} = 2, \gamma^{rk}_{1,g} = 7.8, \gamma^{rk}_{2,g} = 3.7, G = 94$; for Fig. 5b are $s^{rk}_{c} = 9, \gamma^{rk}_{1,g} = 1.25, \gamma^{rk}_{2,g} = 4.9, G = 94$. We note that the statistical characterizations of parameters are again considered as a future work [19].

V. Proposed Individual Preference Probability Generation

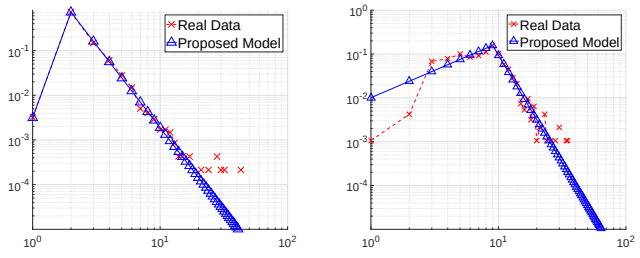
In this section, we first propose an approach that can generate individual preference probabilities of users according to the proposed models. Then the effectiveness of the proposed genreation approach is validated by comparison with real data.

A. Procedure of the Proposed Individual Preference Probability Generation Approach

Here the general procedure of proposed individual preference probability generation is elaborated. To generate the individual preference probabilities of users, we first decide the number of genres in the library and the number of files in each genre, i.e., decide G and $M_g, \forall g$. Then the genre-based conditional popularity distributions $P^{in}_g(.), \forall g$, genre appearance probabilities $P_{ap}(g), \forall g$, and ranking distributions $R_g(.), \forall g$, are generated according to (5), (6), and (7), respectively. Note that these distributions are dataset-specific and are invariant to generating individual preference probabilities of different users.

We next consider to generate the individual preference probabilities of user k. The number of genres in the genre list of user k, i.e., S_k, is first generated according to (2). Then the individual genre popularity distribution $P^{out}_k(.)$ is generated according to (4); the genre list and the specific ranking order of user k is generated according to the proposed ranking order

[5]The characterization of correlations between the appearances of genres in the genre list is considered as future work.



(a) Case 1. (b) Case 2.

Fig. 5: Exemplary comparisons between proposed model and real data of ranking distributions.

generation approach in Alg. 1. The output of the ranking order generation process is the genre index vector $\mathbf{r}_k$ of user k. $\mathbf{r}_k$ contains the indices of genres that appear in the genre list. In addition, the order of the indices in the vector is exactly the ranking order of corresponding genres. Therefore $\mathbf{r}_k$ uniquely specifies the genre list and the ranking order of user k. For example, suppose we have $G = 5$, $S_k = 3$, and $\mathbf{r}_k = [3, 2, 5]$. We know that the genre 2, 3, and 5 are genres in the genre list of user k; genre 3 is the 1st ranked genre; genre 2 is the 2nd ranked genre; and genre 5 is the 3rd ranked genre for user k. For Alg. 1, we provide the following remarks: (I) step 4 is to randomize the filling order of genres at each round; (II) step 7 is to check whether the genre has already been filled into the genre list; (III) step 10 is to check whether the selected genre appears in the genre list and the its ranking value R are less or equal to the size of the genre list; and (IV) step 16 is to generate the final order of genres in the list according to the generated ranking values. For example, suppose $G = 5$, $S_k = 3$, and $\mathbf{r} = [0, 2, 1, 0, 2]$. We would have $\mathbf{r}_k = [3, 2, 5]$ according to step 16 in Alg. 1.

Equipped with genre-based conditional probability distributions $P_g^{in}(.)$ and after generations of individual preference popularity $P_k^{out}(.)$ and genre index vector $\mathbf{r}_k$, individual preference probabilities of user k can then be generated by[6]

$$p_{g,m}^k = f_{k,g}^{out} \times P_g^{in}(m), \tag{8}$$

where

$$f_{k,g}^{out} = \begin{cases} P_k^{out}(i), \text{ entry } i \text{ of } \mathbf{r}_k = g \\ 0, \text{ otherwise} \end{cases}. \tag{9}$$

Eq. (9) indicates that only genres indexed in $\mathbf{r}_k$ have non-zero preference probabilities, and the preference order is given by the order of indices in $\mathbf{r}_k$. For example, suppose that we have $G = 5$, $S_k = 3$, $P_k^{out}(1) = 0.5455$, $P_k^{out}(2) = 0.2727$, $P_k^{out}(3) = 0.1818$, and $\mathbf{r}_k = [3, 2, 5]$. Then $f_{k,1}^{out} = f_{k,4}^{out} = 0$, $f_{k,3}^{out} = P_k^{out}(1) = 0.5455$, $f_{k,2}^{out} = P_k^{out}(2) = 0.2727$, and $f_{k,5}^{out} = P_k^{out}(3) = 0.1818$. We note that, without loss of generality (for the proposed modeling framework), (8) assumes the indices of files within each genre to follow the decending order of the global popularities of files within the genre, i.e., $p_{g,m}^k \geq p_{g,m+1}^k, \forall k$. By combining (8) with (9), the individual preference probabilities $p_{g,m}^k, \forall g, m$ of user k can be obtained. By repeating procedures in this section, individual preference probabilities of different users can be generated.

Algorithm 1 Proposed Ranking Order Genreation Approach

1: **Input:** S_k
2: **Init:** a zero vector $\mathbf{r} = \mathbf{0}$
3: **while** number of non-zero entries in $\mathbf{r} < S_k$ **do**
4: Create a random permutaion vector $\mathbf{P}_v$ with entries being $2, 3, ..., G$ and create an augmented vector $\mathbf{P} = [1|\mathbf{P}_v]$
5: **for** $g = 1 \rightarrow G$ **do**
6: $i = \mathbf{P}(g)$
7: **if** $\mathbf{r}(i) = 0$ **then**
8: $t \backsim binomial(1, P_{ap}(i))$
9: $R \backsim DouZipf(s_{c,i}^{rk}, \gamma_{1,i}^{rk}, \gamma_{2,i}^{rk}, G)$
10: **if** $t = 1$ and $R \leq S_k$ and number of non-zero entries in $\mathbf{r} < S_k$ **then**
11: $\mathbf{r}(i) = R$
12: **end if**
13: **end if**
14: **end for**
15: **end while**
16: $\mathbf{r}_k$ = arrangement of indices of $sort(\mathbf{r}, ascend)$. Break tie by putting the lower index at the lower order. Ignore indices with corresponding values being zero in $\mathbf{r}$.
17: **return** $\mathbf{r}_k$

B. Validations of the Proposed Individual Preference Probability Generation Approach

Here the proposed generation approach is validated by comparing generated results to the underlying real data. To set up the generation approach, parameters of models used by the approach need to be specified. Basically, we use parameters derived from the adopted dataset and consider $G = 94$ and $\sum_{g=1}^{G} M_g = 8996$. The ratios between $M_g, \forall g$ follow the almost identical ratios between the numbers of files of genres in the dataset. This means, for example, suppose we have 1000 files in genre 1 and 2000 files in genre 2 according to real data, and want $M_1 + M_2 = 30$. We would have $M_1 = 10, M_2 = 20$. Parameters for generating $P_{ap}(g)$ and S_k are given by $\gamma_{ap} = 13.5$, $q_{ap} = 100$, and $s_c = 8, \gamma_1 = 1.9, \gamma_2 = 3.8, M_s = 54$, respectively. Since parameters for generating $P_g^{in}(.), \forall g$, $P_k^{out}(.), \forall k$, and $R_g, \forall g$ require certain statistical modeling which is still under development as described in previous sections, to generate the comparison curve, parameters for them are numerically obtained directly from the dataset and used by the generation approach.

The validation of the individual components of the model has been provided in Secs. III and IV. As a validation of the complete model, we investigate whether averaging over the obtained individual user distributions provides the total popularity distribution that was independently extracted from the observed data. Fig. 6 compares the global popularity of files of the dataset with the global popularity of files constructed by

[6]It can be observed that, with the proposed modeling and generator, the file m in genre g is the mth ranked file in the genre-based conditional popularity distribution of genre g. This is because the non-user-specific genre-based conditional popularity distribution is used to approximate the user preferences of files within the genre, and this index arrangement is used for convenience and without loss of the generality.

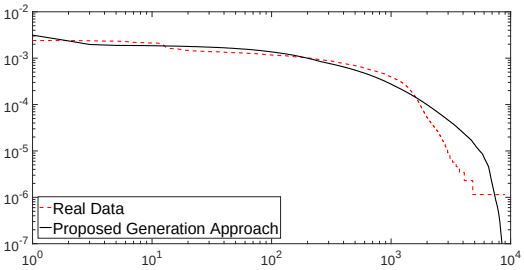


Fig. 6: Comparison between global popularity distributions of files from proposed generation approach and real data.

realizations generated by the proposed approach and shows good agreement. We stress that this only validates that the model can reproduce the data from which it was derived. In future work we will investigate the sensitivity of the model parameters to the dataset, i.e., whether model parameters vary from month to month.

VI. Conclusions and Future Directions

This paper proposed what is to the best of our knowledge the first modeling framework and corresponding statistical models for individual preference probabilities of users for video content based on real-world data. The parameterized model is able to reproduce critical statistics of individual preference. The model is based on, and parameterized by, extensive real-world data sets. The effeciveness of the proposed model and generation approach is validated.

In this work, the statisics of some parameters in the proposed models need to be further investigated and modeled, as do correlations between different patrameters; this, as well as parameterization based on additional data sets, will be done in future work.

Acknowledgements

Part of this work was supported by the National Science Foundation (NSF).

References

[1] N. Golrezaei, A. F. Molisch, A. G. Dimakis, G. Caire, "Femtocaching and device-to-device collaboration: A new architecture for wireless video distribution," *IEEE Commmun. Mag.*, vol. 51, no. 4, pp. 142-149, Apr. 2013.

[2] A. F. Molisch, G. Caire, D. Ott, J. R. Foerster, D. Bethanabhotla, and M. Ji, Caching eliminates the wireless bottleneck in video aware wireless networks, Adv. Elect. Eng., vol. 2014, Nov. 2014, Art. ID 261390.

[3] X. Wang, M. Chen, T. Taleb, A. Ksentini, and V. Leung, "Cache in the air: Exploiting content caching and delivery techniques for 5G systems," *IEEE Commmun. Mag.*, vol. 52, no. 2, pp. 131-139, Feb. 2014.

[4] D. Liu, B. Chen, C. Yang, and A. F. Molisch, "Caching at the wireless edge: Design aspects, challenges, and future directions," *IEEE Commmun. Mag.*, vol. 54, no. 9, pp. 22-28, Sep. 2016.

[5] K. Shanmugam, N. Golrezaei, A. F. Molisch, A. G. Dimakis, G. Caire, "FemtoCaching: Wireless content delivery through distributed caching helpers," *IEEE Trans. Inf. Theory*, vol. 59, no. 12, pp. 8402-8413, Dec. 2013.

[6] A. Liu and V. K. N. Lau, "Exploiting base station caching in MIMO cellular networks: Opportunistic cooperation for video streaming," *IEEE Trans. Sig. Process.*, vol. 63, no. 1, pp. 57-69, Jan. 2015.

[7] Y. Cui, D. Jiang, and Y. Wu, "Analysis and optimization of caching and multicasting in large-scale cache-enabled wireless networks," *IEEE Trans. Wireless Commun.*, vol. 15, no. 7, pp. 5101-5112, Jul. 2016.

[8] N. Golrezaei, P. Mansourifard, A. F. Molisch, and A. G. Dimakis, "Base-Station assisted device-to-device communications for high-throughput wireless video networks," *IEEE Trans. Wireless Commun.*, vol. 13, no. 7, pp. 3665-3676, Jul. 2014.

[9] N. Golrezaei, A. G. Dimakis, and A. F. Molisch "Scaling behavior for device-to-device communications With distributed caching," *IEEE Trans. Inf. Theory*, vol. 60, no. 7, pp. 4286-4298, Jul. 2014.

[10] M. Ji, G. Caire, and A. F. Molisch "Wireless device-to-device caching networks: Basic principles and system performance," *IEEE J. Sel. Area Commun.*, vol. 34, no. 1, pp. 176-189, Jan. 2016.

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

[12] Y. Guo, L. Duan, and R. Zhang, "Cooperative local caching under heterogeneous file preferences," *IEEE Trans. Commun.*, vol. 65, no. 1, pp. 444-457, Jan. 2017.

[13] Y, Pan, C, Pan, H, Zhu, and *et. al.*, "On consideration of content preference and sharing willingness in D2D assisted offloading," *arXiv preprint*, arXiv:1702.00209, Feb. 2017.

[14] M.-C. Lee and A. F. Molisch, "Individual preference aware caching policy design for energy-efficient wireless D2D communications," *IEEE GlOBECOM,* Dec. 2017.

[15] G. Nencioni, N. Sastry, G. Tyson, and *et. al.*, "SCORE: Exploiting global broadcasts to create offline personal channels for on-demand access," *IEEE/ACM Trans. Netw.*, vol. 24, no. 4, pp. 2429-2442, Aug. 2016.

[16] D. Karamshuk, N. Sastry, M. Al-Bassam, A. Secker, and J. Chandaria, "Take-Away TV: Recharging wok commutes with predictive preloading of catch-up TV content," *IEEE J. Sel Commun.*, vol. 34, no. 8, pp. 2091-2101, Aug. 2016.

[17] W. Hoiles, O. N. Gharehshiran, V. Krishnamurthy, N.-D. Dáo, and H. Zhang, "Adaptive caching in the YouTube content distribution network: A revealed preference game-theoretic learning approach," *IEEE Trans. Cogn. Commun. Netw.*, vol. 1, no. 1, pp. 71-84, Mar. 2015.

[18] M. Hefeeda and O. Saleh "Traffic modeling and proportional partial caching for peer-to-peer systems," *IEEE/ACM Trans. Netw.*, vol. 16, no. 6, pp. 1447-1460, Dec. 2008.

[19] M.-C. Lee, A. F. Molisch, N. Sastry, and A. Raman, *IEEE Trans. Wireless Commun.*, to be submitted.