# Characterizing the Rate-Memory Tradeoff in Cache Networks within a Factor of 2

Qian Yu, Student Member, IEEE, Mohammad Ali Maddah-Ali, Member, IEEE, and A. Salman Avestimehr, Senior Member, IEEE

Abstract—We consider a basic caching system, where a single server with a database of N files (e.g. movies) is connected to a set of K users through a shared bottleneck link. Each user has a local cache memory with a size of M files. The system operates in two phases: a placement phase, where each cache memory is populated up to its size from the database, and a following delivery phase, where each user requests a file from the database, and the server is responsible for delivering the requested contents. The objective is to design the two phases to minimize the load (peak or average) of the bottleneck link. We characterize the rate-memory tradeoff of the above caching system within a factor of 2.00884 for both the peak rate and the average rate (under uniform file popularity), improving state of the arts that are within a factor of 4 and 4.7 respectively. Moreover, in a practically important case where the number of files (N) is large, we exactly characterize the tradeoff for systems with no more than 5 users, and characterize the tradeoff within a factor of 2 otherwise. To establish these results, we develop two new converse bounds that improve over the state of the art.

#### I. INTRODUCTION

Caching is a common strategy to mitigate heavy peak-time communication load in a distributed network, via duplicating parts of the content in memories distributed across the network during off-peak times. In other words, caching allows us to trade distributed memory in the network for communication load reduction. Characterizing this fundamental *rate-memory tradeoff* is of great practical interest, and has been a research subject for several decades. For single-cache networks, the *rate-memory tradeoff* has been characterized for various scenarios in the 80s [2]. However, those techniques were found insufficient to tackle the multi-cache cases.

There has been a surge of recent results in information theory that aim at formalizing and characterizing such ratememory tradeoff in multi-cache networks [3]–[17]. In particular, a basic bottleneck caching network was considered in [3], where a set of K users is connected to a server through

Manuscript received February 15, 2017; revised February 07, 2018; accepted Augest 16, 2018. A shorter version of this paper was presented at ISIT, 2017 [1].

Q. Yu and A.S. Avestimehr are with the Department of Electrical Engineering, University of Southern California, Los Angeles, CA, 90089, USA (e-mail: qyu880@usc.edu; avestimehr@ee.usc.edu).

M. A. Maddah-Ali is with Department of Electrical Engineering, Sharif University of Technology, Tehran, 11365, Iran (e-mail: maddah\_ali@sharif.edu).

Communicated by M. Neely, Associate Editor for Shannon Theory.

This work is in part supported by NSF grants CCF-1408639, NETS-1419632, and ONR award N000141612189.

Copyright (c) 2017 IEEE. Personal use of this material is permitted. However, permission to use this material for any other purposes must be obtained from the IEEE by sending a request to pubs-permissions@ieee.org. a shared error-free link. In this setting, each user has a local cache of size M, which can be used to prefetch the contents (a library of N files). The objective is to design the caching functions, such that in a following delivery phase, the server can serve the user demands with efficient bandwidth usage (measured by the communication rate R). For this case, the peak rate vs. memory tradeoff (the tradeoff between maximum R over all possible user demands and M) was formulated and characterized within a factor of 12 [3]. This caching framework has been extended to many scenarios, including decentralized caching [4], online caching [5], caching with nonuniform demands [6]-[8], device-to-device caching [9], caching on file selection networks [10], caching on broadcast channels [11], caching for channels with delayed feedback with channel state information [12], hierarchical cache networks [13], [14], and caching on interference channels [15]-[17], among others. Many of these extensions share similar ideas in terms of the achievability and the converse bounds. Therefore, if we can improve the results for the basic bottleneck caching network, the ideas can be used to improve the results in other cases as well.

In the literature, various approaches have been proposed to improve the bounds on rate-memory tradeoff for the bottleneck network. Several caching schemes have been proposed in [18]–[25], and converse bounds have also been introduced in [10], [26]–[30]. For the case, where the prefetching is uncoded, the exact rate-memory tradeoff for both peak and average rate (under uniform file popularity) and for both centralized and decentralized settings have been established in [24]. However, for the general case, where the cached content can be an arbitrary function of the files in the database, the exact characterization of the tradeoff remains open. In this case, the state of the art is an approximation within a factor of 4 for peak rate [26] and 4.7 for average rate under uniform file popularity [10].

In this paper, we improve the approximation on characterizing the rate-memory tradeoff by proving new informationtheoretic converse bounds, and achieving an approximation within a factor of 2.00884, for both the *peak rate* and the *average rate* under uniform file popularity. These converse bounds hold for the general information theoretic framework, in the sense that there is no constraint on the caching or delivery process. In particular it is not limited to linear coding or uncoded prefetching. This improved characterization is approximately a two-fold improvement with respect to the state of the art in current literature [10], [26].

Furthermore, for a practically important case where the

number of files is large, we exactly characterize the ratememory tradeoff for systems with no more than 5 users. In this case, we also characterize the rate-memory tradeoff within a factor of 2 for networks with an arbitrary number of users, slightly improving our factor-of-2.00884 characterization in the general case. In prior works, despite various attempts, this tradeoff has only been exactly characterized in two instances: the single-user case [3] and, more recently, the two-user case [29].

To prove these results we develop two new converse bounds for cache networks. The first converse is developed based on the idea of enhancing the cutset bound, to effectively capture the isolation of cache contents of the users that belong to the same side of the cut. This approach strictly improves the compound cutset bound, which was used in most of the prior works. Furthermore, using this converse, we are able to characterize both the peak rate and the average rate within factor of 2.00884. To prove this result, we essentially demonstrate that our new converse is within a factor of 2.00884 from the achievable scheme developed in [21] for all possible parameter values.

Moreover, we develop a second converse bound, which is proved by carefully dividing the set of all user demands into certain subsets, and lower bounding the communication rate within each subset separately. Unlike the first converse, it exploits the scenarios where users may have common demands. This enables improvement upon the first converse, and allows exact characterization of the rate-memory tradeoff for systems with up to 5 users.

The rest of this paper is organized as follows. In Section II, we formally define the caching framework and the ratememory tradeoff. Then in Section III we summarize our main results. Section IV proves our first main result, which characterizes the peak rate-memory tradeoff within a constant factor of 2.00884 for all possible parameter values, and characterizes this tradeoff within a factor of 2 when the number of files is large. Section IV proves the converse bound that is needed to establish this characterization. For brevity, we prove the rest of the results in appendices.

#### **II. SYSTEM MODEL AND PROBLEM FORMULATION**

In this section, we formally introduce the system model for the caching problem. Then we define the rate-memory tradeoff for both peak rate and average rate based on the introduced framework, and state the corresponding main problems studied in this paper.

#### A. System Model

We consider a system with one server connected to K users through a shared, error-free link (see Fig. 1). The server has access to a database of N files 1, ..., N, each of size F bits. We assume that the contents of all files, denoted by  $W_1, ..., W_N$ , are i.i.d. random variables, each of which is uniformly distributed on set  $\{1, ..., 2^F\}$ . Each user k has an isolated cache memory of size MF bits, where  $M \in [0, N]$ . For convenience, we define a parameter  $r = \frac{KM}{N}$ .

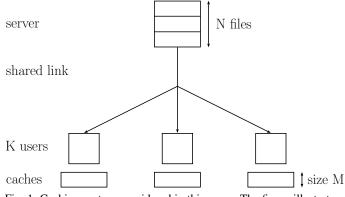


Fig. 1: Caching system considered in this paper. The figure illustrates the case where K = N = 3, and M = 1.

The system operates in two phases: a placement phase and a delivery phase. In the placement phase, the users are given access to the entire database. Each user can fill the contents of their caches using the database without knowledge of their future demands.<sup>1</sup> We denote the cached content of each user k by  $Z_k$ . Then in a following delivery phase, only the server has access to the database of files, and each user requests one of the files in the database. To characterize the requests from the users, we define *demand*  $d = (d_1, ..., d_K)$ , where  $d_k$  is the file requested by user k.

The server is informed of the demand and proceeds by generating a message of size RF bits, denoted by  $X_d$ , as a function of  $W_1, ..., W_N$ , and sends the message over the shared link. R is a fixed real number given the demand d. The quantities RF and R are referred to as the load and the rate of the shared link, respectively. Using the contents  $Z_k$  of its cache and the message  $X_d$  received over the shared link, each user k aims to reconstruct its requested file  $W_{d_k}$ .

#### B. Problem Definition

Based on the above framework, we define the rate-memory tradeoff using the following terminology. We characterize a prefetching scheme by its K caching functions  $\phi = (\phi_1, ..., \phi_K)$ , each of which maps the file contents to the cache content of a specific user:

$$Z_k = \phi_k(W_1, ..., W_N) \qquad \forall k \in \{1, ..., K\}.$$
(1)

Given a prefetching scheme  $\phi$ , we say that a communication rate R is  $\epsilon$ -achievable if and only if, for every request d, there exists a message  $X_d$  of length RF that allows all users to recover their desired file  $d_k$  with a probability of error of at most  $\epsilon$ . Given parameters N, K, and M, we define the minimum peak rate, denoted by  $R^*$ , as the minimum rate that is  $\epsilon$ -achievable over all prefetching schemes for large F and any  $\epsilon > 0$ . Rigorously,

$$R^* = \sup_{\epsilon > 0} \limsup_{F \to \infty} \min_{\phi} \{ R \mid$$

*R* is  $\epsilon$ -achievable given prefetching  $\phi$ } (2)

Similarly for the average rate, we say that a communication rate R is  $\epsilon$ -achievable for demand d, given a prefetching

<sup>&</sup>lt;sup>1</sup>This is due to the fact that in most caching systems the caching phase happens during off-peak hours, in order to improve performance during the peak hours when actual user demands are revealed.

scheme  $\phi$ , if and only if we can create a message  $X_d$  of length RF that allows all users to recover their desired file  $d_k$  with a probability of error of at most  $\epsilon$ . Given parameters N, K, and M, we define the minimum average rate, denoted by  $R_{\text{ave}}^*$ , as the minimum rate over all prefetching schemes such that, we can find a function R(d) that is is  $\epsilon$ -achievable for any demand d, satisfying  $R_{\text{ave}}^* = \mathbb{E}_d[R(d)]$ , where d is uniformly random in  $\mathcal{D} = \{1, ..., N\}^K$ , for large F and any  $\epsilon > 0$ .

Finding the rate-memory tradeoff is essentially finding the values of  $R^*$  and  $R^*_{ave}$  as a function of N, K, and M. In this paper, we aim to find converse bounds that characterize  $R^*$  and  $R^*_{ave}$  within a constant factor. Moreover, we aim to better characterize  $R^*$  and  $R^*_{ave}$  for an important case where N is large, when K and  $\frac{M}{N}$  are fixed.

## C. Related Works

Coded caching was originally proposed in [3], where the peak rate vs. memory tradeoff was characterized within a factor of 12. This result was later extended in [6], where the minimum average rate under uniform file popularity was characterized within a factor of 72. Since then, various efforts has been made on improving these characterizations [10], [26]–[28]. The state of the art is an approximation within a factor of 4 for peak rate [26] and 4.7 for average rate [10].

In this paper, we characterize both the peak rate and the average rate within a factor of 2.00884, which is about a two-fold improvement upon the prior arts. This improvement is achieved by improving both the achievability scheme and the converse. Specifically, we use the achievability scheme we recently proposed in [24] to upper bound the communication rates. This upper bound strictly improves upon the communication rates achieved by [3] (and its relaxed version in [4]), which was relied on by all the above works (i.e., [6], [10], [26]–[28]). It also achieves the exact optimum communication rates among all caching schemes with uncoded prefetching, for all possible values of N, K, and M. As a shorthand notation, we denote the peak and average rates achieved in [24] by  $R_u(N, K, r)$  and  $R_{u,ave}(N, K, r)$ , respectively.<sup>2</sup> More precisely, we define these functions as follows.

**Definition 1.** Given problem parameters N, K, M, and  $r = \frac{KM}{N}$ , we define

$$R_{\rm u}(N,K,r) = \frac{\binom{K}{r+1} - \binom{K-\min\{K,N\}}{r+1}}{\binom{K}{r}},\tag{3}$$

$$R_{\mathrm{u,ave}}(N,K,r) = \mathbb{E}_{\boldsymbol{d}}\left[\frac{\binom{K}{r+1} - \binom{K-N_{\mathrm{e}}(\boldsymbol{d})}{r+1}}{\binom{K}{r}}\right]$$
(4)

for  $r \in \{0, ..., K\}$ , where d is uniformly random in  $\mathcal{D} = \{1, ..., N\}^K$ , and  $N_e(d)$  denotes the number of distinct requests in d.<sup>3</sup> Furthermore, for general (non-integer)  $r \in [0, K]$ ,

 $R_{u}(N, K, r)$  and  $R_{u,ave}(N, K, r)$  are defined as the lower convex envelope of their values at  $r \in \{0, 1, ..., K\}$ , respectively. Specifically, for any non-integer  $r \in [0, K]$ , we have<sup>4</sup>

$$R_{u}(N, K, r) = (r - \lfloor r \rfloor)R_{u}(N, K, \lceil r \rceil) + (\lceil r \rceil - r)R_{u}(N, K, \lfloor r \rfloor),$$

$$R_{u,ave}(N, K, r) =$$
(5)

$$(r - \lfloor r \rfloor)R_{u,ave}(N, K, \lceil r \rceil) + (\lceil r \rceil - r)R_{u,ave}(N, K, \lfloor r \rfloor)$$
(6)

Given the above upper bounds, we develop improved converse bounds in this paper, which provides better characterizations for both the peak rate and the average rate.

#### **III. MAIN RESULTS**

We summarize our main results in the following theorems. **Theorem 1.** For a caching system with K users, a database of N files, and a local cache size of M files at each user, we have

$$\frac{R_{\rm u}(N,K,r)}{2.00884} \le R^* \le R_{\rm u}(N,K,r),\tag{7}$$

$$\frac{R_{u,ave}(N,K,r)}{2.00884} \le R_{ave}^* \le R_{u,ave}(N,K,r).$$
(8)

where  $R_{u}(N, K, r)$  and  $R_{u,ave}(N, K, r)$  are defined in Definition 1. Furthermore, if N is sufficiently large (specifically,  $N \ge \frac{K(K+1)}{2}$ ), we have

$$\frac{R_{\mathfrak{u}}(N,K,r)}{2} \le R^* \le R_{\mathfrak{u}}(N,K,r),\tag{9}$$

$$\frac{R_{u,ave}(N,K,r)}{2} \le R_{ave}^* \le R_{u,ave}(N,K,r).$$
(10)

*Remark* 1. The above theorem characterizes  $R^*$  and  $R^*_{ave}$  within a constant factor of 2.00884 for all possible values of parameters K, N, and M. To the best of our knowledge, this gives the best characterization to date. Prior to this work, the best proved constant factors were 4 for peak rate [26] and 4.7 for average rate (under uniform file popularity) [10]. Furthermore, Theorem 1 characterizes  $R^*$  and  $R^*_{ave}$  for large N within a constant factor of 2.

*Remark* 2. The converse bound that we develop for proving Theorem 1 also immediately results in better approximation of rate-memory tradeoff in other scenarios, such as online caching [5], caching with non-uniform demands [6], and hierarchical caching [14]. For example, in the case of online caching [5], where the current approximation result is within a multiplicative factor of 24, it can be easily shown that this factor can be reduced to 4.01768 using our proposed bounding techniques.

*Remark* 3.  $R_u(N, K, r)$  and  $R_{u,ave}(N, K, r)$ , as defined in Definition 1, are the optimum peak rate and the optimum average rate that can be achieved using uncoded prefetching, as we proved in [24]. This indicates that for the coded caching problem, using uncoded prefetching schemes is within a factor

<sup>&</sup>lt;sup>2</sup>Recall that  $r \triangleq \frac{KM}{N}$ . The letter "u" in the subscript represents "upper bound", and "uncoded prefetching".

<sup>&</sup>lt;sup>3</sup>Here the letter "e" in the subscript represents "effective", given that the function  $N_{e}(d)$  can also be interpreted as the "effective" number of files for any demand d. Specifically, for any demand d, the needed communication rate stated in equation (4) is exactly the peak communication rate stated in equation (3) for a caching system with  $N = N_{e}(d)$  files.

<sup>&</sup>lt;sup>4</sup>Rigorously, the fact that equations (5) and (6) define lower convex envelopes is due to the convexity of  $R_u(N, K, r)$  and  $R_{u,ave}(N, K, r)$ on  $r \in \{0, 1, ..., K\}$ . This convexity was observed in [24] and can be proved using elementary combinatorics. A short proof of the convexity of  $R_u(N, K, r)$  and  $R_{u,ave}(N, K, r)$  can be found in Appendix J.

of 2.00884 optimal for both peak rate and average rate. More interestingly, we can show that even for the improved decentralized scheme we proposed in [24], where each user fills their cache independently without coordination but the delivery scheme was designed to fully exploit the commonality of user demands, the optimum rate is still achieved within a factor of 2.00884 in general, and a factor of 2 for large N.<sup>5</sup>

*Remark* 4. Based on the proof idea of Theorem 1, we can completely characterize the rate-memory tradeoff for the twouser case, for any possible values of N and M, for both peak rate and average rate. Prior to this work, the peak rate vs. memory tradeoff for the two-user case was characterized in [3] for  $N \leq 2$ , and is characterized in [29] for  $N \geq 3$  very recently. However the average rate vs. memory tradeoff has never been completely characterized for any non-trivial case. In this paper, we prove that the exact optimal tradeoff for the average rate for two-user case can be achieved using the caching scheme we provided in [24] (see Appendix H).

To prove the Theorem 1, we derive new converse bounds of  $R^*$  and  $R^*_{ave}$  for all possible values of K, N, and M. We highlight the converse bound of  $R^*$  in the following theorem: **Theorem 2.** For a caching system with K users, a database of N files, and a local cache size of M files at each user,  $R^*$ is lower bounded by

$$R^* \ge s - 1 + \alpha - \frac{s(s-1) - \ell(\ell-1) + 2\alpha s}{2(N - \ell + 1)}M, \quad (11)$$

for any  $s \in \{1, ..., \min\{N, K\}\}$ ,  $\alpha \in [0, 1]$ , where  $\ell \in \{1, ..., s\}$  is the minimum value such that<sup>6</sup>

$$\frac{s(s-1) - \ell(\ell-1)}{2} + \alpha s \le (N - \ell + 1)\ell.$$
(12)

*Remark* 5. The above theorem improves the state of the art in various scenarios. For example, when N is sufficiently large (i.e.,  $N \ge \frac{K(K+1)}{2}$ ), the above theorem gives tight converse bound for  $\frac{KM}{N} \le 1$ , as shown in (23). The above matching converse can not be proved directly using converse bounds provided in [10], [26]–[30] (e.g., for K = 4, N = 10, and M = 1, none of these bounds give  $R^* \ge 3$ ).

*Remark* 6. Although Theorem 2 gives infinitely many linear converse bounds on  $R^*$ , the region of the memory-rate pair  $(M, R^*)$  characterized by Theorem 2 has a simple shape with finite corner points. Specifically, by applying the arguments used in the proof of Theorem 1, one can show that the exact bounded region given by Theorem 2 is bounded by the lower convex envelop of points  $\{(\frac{N-\ell+1}{s}, \frac{s-1}{2} + \frac{\ell(\ell-1)}{2s}) \mid s \in \{1, ..., J\}, \ell \in \{1, ..., s\}\} \cup \{(0, J)\}$ , where  $J = \min\{N, K\}$ .

For the case of large N, we can exactly characterize the values of  $R^*$  and  $R^*_{ave}$  for  $K \leq 5$ . We formally state this result in the following theorem:

**Theorem 3.** For a caching system with K users, a database of N files, and a local cache size of M files at each user, we

<sup>6</sup>Such  $\ell$  always exists, because when  $\ell = s$ , (12) can be written as  $\alpha s \leq (N - s + 1)s$ , which always holds true.

have

$$R^* = R^*_{\text{ave}} = R_u(N, K, r)$$
(13)

for large N (i.e.,  $N \to +\infty$ ) when  $K \leq 5$ , where  $R_u(N, K, r)$  is defined in Definition 1.<sup>7</sup>

*Remark* 7. As discussed in [4], the special case of large N is important to handle asynchronous demands. More specifically, [4] showed that asynchronous demands can be handled by splitting each file into many subfiles, and delivering concurrent subfile requests using the optimum caching schemes. In this case, we essentially need to solve the caching problem when the number of files (i.e., the subfiles) is large, but the fraction of files that can be stored at each user is fixed. In this paper, we completely characterize this tradeoff for systems with up to 5 users, for both peak rate and average rate, while in prior works, this tradeoff has only been exactly characterized in two instances: the single-user case [3] and, more recently, the two-user case [29].

*Remark* 8. Although Theorem 3 only consider systems with up to 5 users, the converse bounds used in its proof also tightly characterize the minimum communication rate in many cases even for systems with more than 5 users. For both peak rate and average rate, we can show that more than half of the convex envelope achieved by [24] are optimal for large N (e.g., see Lemma 4 for peak rate).

To prove Theorem 3, we state the following Theorem, which provides tighter converse bounds on  $R^*$  for certain values of N, K, and M.

**Theorem 4.** For a caching system with K users, a database of N files, and a local cache size of M files at each user,  $R^*$  is lower bounded by

$$R^* \ge \begin{cases} \frac{2K - n + 1}{n + 1} - \frac{K(K+1)}{n(n+1)} \cdot \frac{M}{N} & \text{if } \beta + \alpha \frac{K - 2n - 1}{2} \le 0, \\\\ \frac{2K - n + 1}{n + 1} - \frac{2K(K-n)}{n(n+1)} \cdot \frac{M}{N - \beta} & \text{otherwise}, \end{cases}$$
(14)

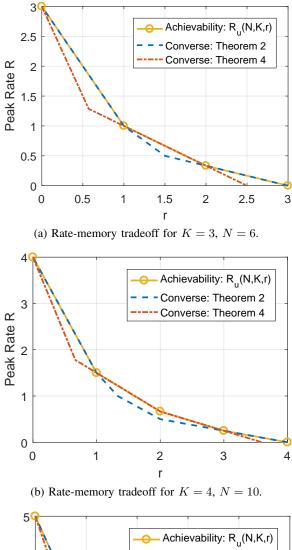
for any  $n \in \{\max\{1, K - N + 1\}, ..., K - 1\}$ , where  $\alpha = \lfloor \frac{N-1}{K-n} \rfloor$  and  $\beta = N - \alpha(K - n)$ .

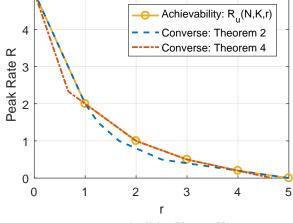
*Remark* 9. The above theorem improves Theorem 2 and the state of the art in many cases. For example, when  $r \in \left[\left[K-1-\frac{N-1}{\left\lceil\frac{2N}{K+1}\right\rceil}\right], K-1\right)$ , the converse bound (14) given by  $n = \lfloor r+1 \rfloor$  is tight and we have  $R^* = R_u(N, K, r)$ . This result can not be proved in general using the converse bounds provided in [10], [26]–[30] (e.g., for K = 4, N = 10, and M = 4, none of these bounds give  $R^* \geq 1$ ).

*Remark* 10. We numerically compare our two converse bounds (i.e., Theorem 2 and Theorem 4), benchmarked against the upper bound  $R_u(N, K, r)$  we achieved in [24] under three different settings (see Fig. 2). In all these cases, the two converse bounds together provide a tight characterization: Theorem 2 is tight for  $r \le 1$  and  $r \ge K - 1$ , and Theorem 4 is tight for  $1 \le r \le K - 1$ . The same holds true in the proof of Theorem 3, where the number of users is no more than 5 but the number of files is large.

<sup>7</sup>Rigorously, we show that the maximum possible gap between  $R^*$ ,  $R^*_{ave}$ , and  $R_u(N, K, r)$  over  $M \in [0, N]$  approaches 0 as N goes to infinity.

<sup>&</sup>lt;sup>5</sup>This can be proved based on the fact that, in the proof of Theorem 1, we showed the communication rates of the decentralized caching scheme we proposed in [24] (e.g.,  $R_{dec}(M)$  for the peak rate) are within constant factor optimal as intermediate steps.





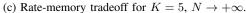


Fig. 2: Numerical comparison among the two converse bounds presented in Theorem 2 and Theorem 4, and the upper bound achieved in [24]. Our converse bounds tightly characterize the peak rate-memory tradeoff in all three presented scenarios.

In the rest of this paper, we prove Theorem 1 for the peak rate in Section IV, and we prove Theorem 2 in Section V. For brevity, we prove the rest of the results in the appendices. Specifically, Appendix A proves Theorem 3 for the peak rate, Appendix B proves Theorem 4, Appendix G proves Theorem 1 for the average rate, and Appendix I proves Theorem 3 for the average rate.

## IV. PROOF OF THEOREM 1 FOR PEAK RATE

In this section, we prove Theorem 1 assuming the correctness of Theorem 2. The proof of Theorem 2 can be found in Section V. For brevity, we only prove Theorem 1 for the peak rate (i.e., inequalities (7) and (9)) within this section. The proof for the average rate (i.e., inequalities (8) and (10)) can be found in Appendix G.

We start by proving the general factor-of-2.00884 characterization for inequality (7). Then we focus on the special case of  $N \ge \frac{K(K+1)}{2}$  and prove inequality (9). As mentioned in Remark 3, the upper bounds of  $R^*$  stated in Theorem 1 can be proved using the caching scheme provided in [24]. Hence, it suffices to prove the lower bounds of (7) and (9).

#### A. Proof of inequality (7)

The proof of inequality (7) consists of 2 steps. In Step 1, we first prove, assuming the correctness of Theorem 2, that the memory-rate pair  $(M, R^*)$  is lower bounded by the lower convex envelope of a set of points in  $S_{\text{Lower}} \cup \{(0, J)\}$ , where

$$S_{\text{Lower}} = \left\{ (M, R) = \left( \frac{N - \ell + 1}{s}, \frac{s - 1}{2} + \frac{\ell(\ell - 1)}{2s} \right) \right|$$
$$s \in \{1, ..., J\}, \ell \in \{1, ..., s\} \right\}$$
(15)

where  $J = \min\{N, K\}$ , given parameters N and K. Then in Step 2, we exploit the convexity of the upper bound  $R_u(N, K, r)$ , and prove that it is within a factor of 2.00884 from the above converse by checking all the corner points of the envelope.

For Step 1, we first prove that  $R^*$  is lower bounded by the convex envelope. To prove this statement, it is sufficient to show that any linear function that lower bounds all points in  $S_{\text{Lower}} \cup \{(0, J)\}$ , also lower bounds the point  $(M, R^*)$ . We prove this for any such linear function, denoted by A + BM, by first finding a converse bound of  $R^*$  using Theorem 2 with certain parameters s and  $\alpha$ , and then proving that this converse bound is lower bounded by the linear function. We consider the following 2 possible cases:

If  $A \ge 0$ , note that (0, J) should be lower bounded by the linear function, so we have  $A \le J$ . Thus, we can choose  $s = \lceil A \rceil$ ,  $\alpha = A - s + 1$ , and let  $\ell$  be the minimum value in  $\{1, ..., s\}$  such that (12) holds. Because  $\left(\frac{N-\ell+1}{s}, \frac{s-1}{2} + \frac{\ell(\ell-1)}{2s}\right) \in \mathcal{S}_{\text{Lower}}$ , we have

$$A + B\frac{N - \ell + 1}{s} \le \frac{s - 1}{2} + \frac{\ell(\ell - 1)}{2s}.$$
 (16)

By the definition of  $\alpha$ , we have  $A = s - 1 + \alpha$ . Consequently, the slope B can be upper bounded as follows:

$$B \leq \frac{s(s-1) + \ell(\ell-1) - 2As}{2(N-\ell+1)} = -\frac{s(s-1) - \ell(\ell-1) + 2\alpha s}{2(N-\ell+1)}.$$
(17)

Thus, for any  $M \ge 0$ , we have

$$A + BM \le s - 1 + \alpha - \frac{s(s-1) - \ell(\ell-1) + 2\alpha s}{2(N - \ell + 1)}M.$$
(18)

Note that the RHS of the above inequality is exactly the lower bound provided in Theorem 2. Hence,  $A + BM \le R^*$ .

If A < 0, let  $s = \ell = 1$ , we have  $(N, 0) \in \mathcal{S}_{\text{Lower}}$  from (15). Hence,  $A + BN \leq 0$ , and for any  $M \in [0, N]$  we have A(N - M) + (A + BN)M

$$A + BM = \frac{A(N - M) + (A + BN)M}{N} \le 0.$$
 (19)

Obviously  $R^* \ge 0$ , hence we have  $A + BM \le R^*$ .

Combining the above two cases, we have proved that the memory-rate pair  $(M, R^*)$  is lower bounded by the lower convex envelope of  $S_{\text{Lower}} \cup \{(0, J)\}$ . This completes the proof of Step 1.

For Step 2, we only need to prove that the ratio of  $R_u(N, K, r)$  to the lower convex envelope of  $S_{Lower} \cup \{(0, J)\}$  is at most 2.00884. As mentioned at the beginning of this proof, given that the upper bound  $R_u(N, K, r)$  is convex,<sup>8</sup> this ratio can only be maximized at the corner points of the envelope, which is a subset of  $S_{Lower} \cup \{(0, J)\}$ . Hence, we only need to check that  $R_u(N, K, r) \leq 2.00884R$  holds for any  $(M, R) \in S_{Lower} \cup \{(0, J)\}$ .

To further simplify the problem, we upper bound  $R_{\rm u}(N,K,r)$  using the following inequality, which can be easily proved using the results of [24]:<sup>9</sup>

$$R_{\rm u}(N,K,r) \le R_{\rm dec}(M) \triangleq \frac{N-M}{M} (1 - (1 - \frac{M}{N})^J).$$
 (20)

Consequently, to prove inequality (7), it suffices to prove the following lemma.

**Lemma 1.** For any  $(M, R) \in S_{\text{Lower}} \cup \{(0, J)\}$ , we have  $R_{\text{dec}}(M) \leq 2.00884R$ .

The proof of Lemma 1 can be found in Appendix C. Assuming its correctness, we have  $R_u(N, K, r) \leq 2.00884R^*$ for all possible parameter values of N, K, and M. This completes the proof of inequality (7).

### B. Proof of inequality (9)

Now we prove that  $R^* \geq \frac{R_u(N,K,r)}{2}$  holds for any  $N \geq \frac{K(K+1)}{2}$ . In this case, we can verify that inequality (12) holds for any  $s \in \{1, ..., K\}$ ,  $\alpha = 1$ , and  $\ell = 1$ . Consequently, from Theorem 2,  $R^*$  can be bounded as follows:

$$R^* \ge s - 1 + 1 - \frac{s(s-1) + 2s}{2(N-1+1)}M$$
$$= s - \frac{s^2 + s}{2} \cdot \frac{M}{N}.$$
(21)

Then we prove  $R^* \geq \frac{R_u(N,K,r)}{2}$  by considering the following 2 possible cases: If  $\frac{KM}{N} \leq 1$ , we have

$$R_{\rm u}(N,K,r) = K - \frac{K^2 + K}{2} \cdot \frac{M}{N}$$
 (22)

as defined in Definition 1. Let s = K, we have the following bounds from (21) which tightly characterizes  $R_u(N, K, r)$ :

$$R^* \ge K - \frac{K^2 + K}{2} \cdot \frac{M}{N} = R_{u}(N, K, r) \ge \frac{R_{u}(N, K, r)}{2}.$$
(23)

<sup>8</sup>A short proof can be found in Appendix J

<sup>9</sup>Here the upper bound  $R_{dec}(M)$  is the exact minimum communication rate needed for decentralized caching with uncoded prefetching, as proved in [24]. When M = 0,  $R_{dec}(M) \triangleq J$ . If  $\frac{KM}{N} > 1$ , let  $s = \lfloor \frac{N}{M} \rfloor$ , we have  $\frac{M}{N} \in [\frac{1}{s+1}, \frac{1}{s}]$ . Consequently, we can derive the following lower bound on  $R^*$ :

$$R^* \ge s - \frac{s^2 + s}{2} \cdot \frac{M}{N}$$

$$= \frac{N - M}{2M} + \frac{s^2 + s}{2} \cdot \frac{N}{M} \cdot \left(\frac{M}{N} - \frac{1}{s+1}\right) \cdot \left(\frac{1}{s} - \frac{M}{N}\right)$$

$$\ge \frac{N - M}{2M}.$$
(24)

As mentioned earlier in this section, the following upper bound can be easily proved using the results of [24]:

$$R_{\rm u}(N,K,r) \le \frac{N-M}{M} (1 - (1 - \frac{M}{N})^K).$$
 (25)

Consequently, we have  $R_u(N, K, r) \leq \frac{N-M}{M} \leq 2R^*$ . To conclude, we have proved  $R^* \geq \frac{R_u(N, K, r)}{2}$  for both cases. Hence, inequality (9) holds for large N for any possible values of K and M.

#### V. PROOF OF THEOREM 2

Before proving the converse bound stated in Theorem 2, we first present the following key lemma, which gives a lower bound on any  $\epsilon$ -achievable rate given any prefetching scheme. **Lemma 2.** Consider a coded caching problem with parameters N and K. Given a certain prefetching scheme, for any demand d, any  $\epsilon$ -achievable rate R is lower bounded by <sup>10</sup>

$$R \ge \frac{1}{F} \left( \sum_{k=1}^{\min\{N,K\}} H(W_{d_k} | Z_{\{1,\dots,k\}}, W_{\{d_1,\dots,d_{k-1}\}}) \right) - \min\{N,K\} (\frac{1}{F} + \epsilon).$$
(26)

The above lemma is developed based on the idea of enhancing the cutset bound, which is further explained in the proof of this lemma in Appendix D. One can show that this approach strictly improves the compound cutset bound, which was used in most of the prior works. We now continue to prove Theorem 2 assuming the correctness of Lemma 2.

The rest of the proof consists of two steps. In Step 1, we exploit the homogeneity of the problem, and derive a symmetrized version of the converse presented in Lemma 2. Then in Step 2, we derive the converse bound in Theorem 2, which is independent of the prefetching scheme, by essentially minimize the symmetrized converse over all possible designs.

For Step 1, we observe that the caching problem proposed in this paper assumes that all users has the same cache size, and all files are of the same size. To fully utilize this homogeneity, we define the following useful notations. For any positive integer *i*, we denote the set of all permutations of  $\{1, ..., i\}$  by  $\mathcal{P}_i$ . For any set  $S \subseteq \{1, ..., i\}$  and any permutation  $p \in \mathcal{P}_i$ , we define  $pS = \{p(s) \mid s \in S\}$ . For any subsets  $\mathcal{A} \subseteq \{1, ..., N\}$ and  $\mathcal{B} \subseteq \{1, ..., K\}$ , we define

$$H^*(W_{\mathcal{A}}, Z_{\mathcal{B}}) \triangleq \frac{1}{N!K!} \sum_{p \in \mathcal{P}_N, q \in \mathcal{P}_K} H(W_{p\mathcal{A}}, Z_{q\mathcal{B}}).$$
(27)

<sup>10</sup>By an abuse of notation, we denote a sub-array by using a set of indices as the subscript. Besides, we define  $\{d_1, ..., d_{k-1}\} = \emptyset$  for k = 1. Similar convention will be used throughout this paper.

Similarly, we define the same notation for conditional entropy in the same way. We can verify that the functions defined above satisfies all Shannon's inequalities. I.e., for any sets of random variables  $\mathcal{A}$ ,  $\mathcal{B}$  and  $\mathcal{C}$ , we have

$$H^*(\mathcal{A}|\mathcal{B}) \ge H^*(\mathcal{A}|\mathcal{B}, \mathcal{C}).$$
(28)

Note that from the homogeneity of the problem, for any  $\epsilon$ -achievable rate R, Lemma 2 holds for any demands, under any possible relabeling of the users. Thus, by considering the class of demands where at least min $\{N, K\}$  files are requested, we have

$$R \ge \frac{1}{F} \left( \sum_{k=1}^{\min\{N,K\}} H(W_{q(k)} | Z_{\{p(1),\dots,p(k)\}}, W_{\{q(1),\dots,q(k-1)\}}) \right) - \min\{N,K\} \left(\frac{1}{F} + \epsilon\right)$$
(29)

for any  $p \in \mathcal{P}_K$  and  $q \in \mathcal{P}_N$ . Averaging the above bound over all possible p and q, we have

$$R \ge \frac{1}{F} \left( \sum_{k=1}^{\min\{N,K\}} H^*(W_k | Z_{\{1,\dots,k\}}, W_{\{1,\dots,k-1\}}) \right) - \min\{N,K\} \left(\frac{1}{F} + \epsilon\right).$$
(30)

Recall that  $R^*$  is defined to be the minimum  $\epsilon$ -achievable rate over all prefetching scheme  $\phi$  for large F for any  $\epsilon > 0$ , we have

$$R^{*} \geq \sup_{\epsilon > 0} \limsup_{F \to \infty} \min_{\phi} \{\frac{1}{F} (\sum_{k=1}^{\min\{N,K\}} H^{*}(W_{k}|Z_{\{1,...,k\}}, W_{\{1,...,k-1\}})) - \min\{N,K\} (\frac{1}{F} + \epsilon)\} = \sup_{\epsilon > 0} \limsup_{F \to \infty} \min_{\phi} \{\frac{1}{F} (\sum_{k=1}^{\min\{N,K\}} H^{*}(W_{k}|Z_{\{1,...,k\}}, W_{\{1,...,k-1\}}))\} \geq \inf_{F \in \mathbb{N}_{+}} \min_{\phi} \{\frac{1}{F} (\sum_{k=1}^{\min\{N,K\}} H^{*}(W_{k}|Z_{\{1,...,k\}}, W_{\{1,...,k-1\}}))\}. (31)$$

Now we have derived a symmetrized version of the converse bound. To simplify the discussion, we define  $R_{A}(F, \phi) = \frac{1}{F} \sum_{k=1}^{\min\{N,K\}} H^{*}(W_{k}|Z_{\{1,...,k\}}, W_{\{1,...,k-1\}})$ . Consequently,  $R^{*} \geq \inf_{F \in \mathbb{N}_{+}} \min_{\phi} R_{A}(F, \phi).$ (32)

For Step 2, as mentioned previously in this proof, to derive the converse bound presented in Theorem 2, we aim to minimize the symmetrized converse  $R_A(F, \phi)$  over all prefetching scheme  $\phi$ . Moreover, we need to prove that it is no less than the RHS of (11) for any parameters s and  $\alpha$ . We present the following lemma, which essentially solves this problem.

**Lemma 3.** For any parameters  $s \in \{1, ..., \min\{N, K\}\}, \alpha \in$ 

[0,1], and any prefetching scheme  $\phi$ , we have

$$R_{\rm A}(F,\phi) \ge s - 1 + \alpha - \frac{s(s-1) - \ell(\ell-1) + 2\alpha s}{2(N-\ell+1)}M,$$
(33)

where 
$$\ell \in \{1, ..., s\}$$
 is the minimum value such that  

$$\frac{s(s-1) - \ell(\ell-1)}{2} + \alpha s \le (N - \ell + 1)\ell.$$
(34)

The proof of Lemma 3 can be found in Appendix E. Note that the lower bound in the above lemma is identical to the converse in Theorem 2. Assuming its correctness, then given any s and  $\alpha$ , we can bound  $R^*$  as follows:

$$R^* \ge \inf_{F \in \mathbb{N}_+} \min_{\phi} R^*_{A}(F,\phi)$$
  
$$\ge (s-1+\alpha) - \frac{s(s-1) - \ell(\ell-1) + 2\alpha s}{2(N-\ell+1)} M. \quad (35)$$

This completes the proof of Theorem 2.

## VI. CONCLUSION

In this paper, we developed novel converse bounding techniques for caching networks, and characterized the ratememory tradeoff of the basic bottleneck caching network within a factor of 2.00884 for both the peak rate and the average rate. This is approximately a two-fold improvement with respect to the state of the art. We also provided tight characterization of rate-memory tradeoff for systems with no more than 5 users, when the number of files is large. The results of this paper can also be used to improve the approximation of rate-memory tradeoff in several other settings, such as online caching, caching with non-uniform demands, and hierarchical caching.

#### APPENDIX A PROOF OF THEOREM 3 FOR PEAK RATE

In this section, we prove Theorem 3 assuming the correctness of Theorem 4. The proof of Theorem 4 can be found in Appendix B. For brevity, we only prove Theorem 3 for the peak rate (i.e.,  $R^* = R_u(N, K, r)$  for large N) within this section. The proof for the average rate (i.e.,  $R^*_{ave} = R_u(N, K, r)$ for large N) can be found in Appendix G.

As mentioned previously, the rate  $R_u(N, K, r)$  can be exactly achieved using the caching scheme proposed in [24]. Hence, to prove Theorem 3, it is sufficient to show that  $R^* \ge R_u(N, K, r)$  for large N (i.e.,  $N \to +\infty$ ) when  $K \le 5$ . This statement can be easily proved using the following lemma:

**Lemma 4.** For a caching problem with parameters K, N, and M, we have  $R^* \ge R_u(N, K, r)$  for large N, if  $r \le 1$  or  $r \ge \lceil \frac{K-3}{2} \rceil$ .

Assuming the correctness of Lemma 4, and noting that the condition in Lemma 4 (i.e.,  $r \leq 1$  or  $r \geq \lceil \frac{K-3}{2} \rceil$ ) always holds true for  $K \leq 5$ , we have  $R^* \geq R_u(N, K, r)$  for large N and for all possible values of M, in any caching system with no more than 5 users. Hence, to prove Theorem 3, it suffices to prove Lemma 4. We prove this lemma as follows, using Theorem 2 and Theorem 4.

*Proof of Lemma 4.* We start by focusing on two easier cases,  $r \leq 1$  and  $r \geq K - 1$ . When  $r \leq 1$ , the inequality  $R^* \geq r$ 

 $R_{\rm u}(N, K, r)$  is already proved in Section IV and given by (23), for  $N \geq \frac{K(K+1)}{2}$ . When  $r \geq K-1$ , we have  $R^* \geq 1 - \frac{M}{N} = R_{\rm u}(N, K, r)$ , which can be proved by choosing s = 1 and  $\alpha = 1$  for Theorem 2. Hence, we only need to focus on the case where  $r \in \left[\max\{\lceil \frac{K-3}{2} \rceil, 1\}, K-1\right)$ , and show that for large N, the maximum possible gap between  $R^*$  and  $R_{\rm u}(N, K, r)$  approaches 0.

We prove this result using Theorem 4. Essentially, we need to find parameter  $n \in \{1, ..., K-1\}$  for Theorem 4, such that the corresponding converse bound approaches  $R_u(N, K, r)$  for large N.

Let n = |r+1|, we have

$$R_{\rm u}(N,K,r) = \frac{2K-n+1}{n+1} - \frac{K(K+1)}{n(n+1)} \cdot \frac{M}{N}$$
(36)

by definition, for sufficiently large N (more specifically,  $N \ge K - n + 1$ ). Under the same condition for large N, we have  $n \in \{\max\{1, K - N + 1\}, ..., K - 1\}$  given  $r \in [1, K - 1)$ . Hence, we can use n as the parameter of Theorem 4. Now we prove the tightness of this converse bound by considering the following two possible cases:

If  $n > \frac{K-1}{2}$ , we have K - 2n - 1 < 0. Recall that  $\alpha = \lfloor \frac{N-1}{K-n} \rfloor$  and  $\beta = N - \alpha(K - n)$ . We can prove that when N is sufficiently large (i.e.  $N \ge \frac{2(K-n)^2}{2n+1-K} + 1$ ), the condition  $\beta + \alpha \frac{K-2n-1}{2} \le 0$  is always satisfied. Consequently,

$$R^* \ge \frac{2K - n + 1}{n + 1} - \frac{K(K + 1)}{n(n + 1)} \cdot \frac{M}{N} = R_{u}(N, K, r).$$
(37)

If  $n \leq \frac{K-1}{2}$ , because we are considering the case where  $r \geq \lceil \frac{K-3}{2} \rceil$ , we have  $n = \frac{K-1}{2}$ . Hence, we can verify that  $\beta + \alpha \frac{K-2n-1}{2} \leq 0$  does not hold for any N. Consequently,

$$R^* \ge \frac{2K - n + 1}{n + 1} - \frac{2K(K - n)}{n(n + 1)} \cdot \frac{M}{N - \beta}$$
$$= \frac{2K - n + 1}{n + 1} - \frac{K(K + 1)}{n(n + 1)} \cdot \frac{M}{N - \beta}.$$
(38)

As N approaches infinity,  $\beta$  is upper bounded by a constant. Hence, we have  $\lim_{N \to +\infty} \frac{N}{N-\beta} = 1$ . Therefore, from (36) and (38), we have

$$\lim_{N \to +\infty} (R^* - R_{u}(N, K, r)) \geq \lim_{N \to +\infty} \frac{K(K+1)}{n(n+1)} \cdot \left(\frac{M}{N} - \frac{M}{N-\beta}\right)$$
$$= \lim_{N \to +\infty} \frac{r(K+1)}{n(n+1)} \cdot \left(1 - \frac{N}{N-\beta}\right)$$
$$= 0. \tag{39}$$

## APPENDIX B Proof of Theorem 4

Before proving the converse bounds stated in Theorem 4, we first present the following key lemma, which gives a lower bound on any  $\epsilon$ -achievable rate given any prefetching scheme.

**Lemma 5.** Consider a coded caching problem with parameters N and K. Given a certain prefetching scheme, any  $\epsilon$ -

achievable rate R is lower bounded by <sup>11</sup> RF  $\geq H^*(W_1|Z_1)$ 

$$+\frac{2}{n(n+1)\alpha}\left(\alpha n(K-n)F - nH^{*}(Z_{1}|W_{\{1,...,\beta\}}) - \sum_{i=0}^{K-n-1} H^{*}(Z_{1}|W_{\{1,...,\beta+i\alpha\}})\right) - \frac{2K-n+1}{n+1}(1+\epsilon F)$$
(40)

for any integer  $n \in \{\max\{1, K - N + 1\}, ..., K - 1\}$ , where  $\alpha = \lfloor \frac{N-1}{K-n} \rfloor$  and  $\beta = N - \alpha(K - n)$ .

We postpone the proof of the above lemma to Appendix F, and continue to prove Theorem 4 assuming its correctness. To simplify the discussion, we define

$$R_{\rm B}(F,\phi) = \frac{1}{F} \left( H^*(W_1|Z_1) + \frac{2}{n(n+1)\alpha} \left( \alpha n(K-n)F - nH^*(Z_1|W_{\{1,\dots,\beta\}}) - \sum_{i=0}^{K-n-1} H^*(Z_1|W_{\{1,\dots,\beta+i\alpha\}}) \right) \right).$$
(41)

Using Lemma 5, we have

$$R \ge R_{\rm B}(F, \phi) - \frac{2K - n + 1}{n + 1} (\frac{1}{F} + \epsilon)$$
 (42)

if R is  $\epsilon$ -achievable. Recall that  $R^*$  is defined to be the minimum  $\epsilon$ -achievable rate over all prefetching scheme  $\phi$  for large F for any  $\epsilon > 0$ , we have the following lower bound on  $R^*$ :

$$R^* \geq \sup_{\epsilon>0} \limsup_{F \to \infty} \min_{\phi} \{R_{\mathrm{B}}(F,\phi) - \frac{2K - n + 1}{n + 1} (\frac{1}{F} + \epsilon)\}$$
  
= 
$$\sup_{\epsilon>0} \limsup_{F \to \infty} \min_{\phi} R_{\mathrm{B}}(F,\phi)$$
  
$$\geq \inf_{E \in \mathbb{N}} \min_{\phi} R_{\mathrm{B}}(F,\phi).$$
(43)

Hence, to prove Theorem 4, we only need to prove that for any prefetching scheme  $\phi$ ,  $R_{\rm B}(F, \phi)$  is lower bounded by the converse bounds given in Theorem 4 for any valid parameter n.

Now consider any  $n \in {\max\{1, K - N + 1\}, ..., K - 1\}}$ . For brevity, we define

$$\theta = \left(K\beta + \frac{(K-n)(K-n-1)}{2}\alpha\right).$$
 (44)

Equivalently, we have

$$\theta = n\beta + \sum_{i=0}^{K-n-1} (\beta + i\alpha).$$
(45)

Hence,

<sup>11</sup>Here we adopt the notation of  $H^*(W_A, Z_B)$  which is defined in the proof of Theorem 2.

$$\theta H^{*}(W_{1}|Z_{1}) \geq nH^{*}(W_{\{1,...,\beta\}}|Z_{1}) + \sum_{i=0}^{K-n-1} H^{*}(W_{\{1,...,\beta+i\alpha\}}|Z_{1}) = \theta F + nH^{*}(Z_{1}|W_{\{1,...,\beta\}}) + \sum_{i=0}^{K-n-1} H^{*}(Z_{1}|W_{\{1,...,\beta+i\alpha\}}) - KH^{*}(Z_{1}).$$

$$(46)$$

From (41) and (46), we have

$$R_{\rm B}(F, \phi)F \ge \left(1 - \frac{2\theta}{n(n+1)\alpha}\right)H^*(W_1|Z_1) + \frac{2}{n(n+1)\alpha}(\theta F - KH^*(Z_1) + \alpha n(K-n)F).$$
(47)

Depending on the value of  $\theta$ , we bound  $H^*(W_1|Z_1)$  in 2 different ways:

When  $1 \ge \frac{2\theta}{n(n+1)\alpha}$ , this is exactly the case where  $\beta$  +  $\alpha \frac{K-2n-1}{2} \leq 0$  holds. We use the following bound:

$$H^*(W_1|Z_1) \ge F - \frac{H^*(Z_1)}{N}.$$
 (48)

Consequently,

$$R_{\rm B}(F,\phi)F \ge \left(1 - \frac{2\theta}{n(n+1)\alpha}\right) \left(F - \frac{H^*(Z_1)}{N}\right) \\ + \frac{2}{n(n+1)\alpha}(\theta F - KH^*(Z_1) \\ + \alpha n(K-n)F).$$
(49)

Given  $\theta$  defined in (44), and  $\beta = N - \alpha(K - n)$  as defined Since  $s \ge \ell$ , we have  $\frac{\ell - 1}{\ell} \le \frac{s - 1}{s}$ . Consequently,

$$R_{\rm B}(F,\phi)F = \frac{2K-n+1}{n+1}F - \frac{K(K+1)}{n(n+1)} \cdot \frac{H^*(Z_1)}{N}$$
$$\geq \frac{2K-n+1}{n+1}F - \frac{K(K+1)}{n(n+1)} \cdot \frac{M}{N}F.$$
 (50)

Hence we have the follows from (43):

$$R^* \ge \frac{2K - n + 1}{n + 1} - \frac{K(K + 1)}{n(n + 1)} \cdot \frac{M}{N}.$$
 (51)

On the other hand, when  $1 < \frac{2\theta}{n(n+1)\alpha}$ , this is exactly the case where  $\beta + \alpha \frac{K-2n-1}{2} \leq 0$  does not hold. We use  $H^*(W_1|Z_1) \leq F$ . Similarly,

$$R_{\rm B}(F,\phi)F \ge \left(1 - \frac{2\theta}{n(n+1)\alpha}\right)F \\ + \frac{2}{n(n+1)\alpha}(\theta F - KH^*(Z_1)) \\ + \alpha n(K-n)F) \\ = \frac{2K - n + 1}{n+1}F - \frac{2K(K-n)}{n(n+1)} \cdot \frac{H^*(Z_1)}{N - \beta} \\ \ge \frac{2K - n + 1}{n+1}F - \frac{2K(K-n)}{n(n+1)} \cdot \frac{M}{N - \beta}F.$$
(52)

Hence,

$$R^* \ge \inf_{F \in \mathbb{N}_+} \min_{\phi} R_{\mathbf{B}}(F, \phi)$$

$$\geq \frac{2K-n+1}{n+1} - \frac{2K(K-n)}{n(n+1)} \cdot \frac{M}{N-\beta}.$$
 (53)

To conclude, we have proved that the converse bound given in Theorem 4 holds for any valid parameter n.

## APPENDIX C **PROOF OF LEMMA 1**

In this appendix, we aim to prove that for any  $(M, R) \in$  $\mathcal{S}_{\text{Lower}} \cup \{(0, J)\}, R_{\text{dec}}(M) \leq 2.00884R.$  Note that if (M, R) = (0, J), we have  $R_{dec}(M) = J \le 2.00884R$ . Hence, it suffices to consider the case where  $(M, R) \in \mathcal{S}_{\text{Lower}}$ .

In this case, we can find  $s \in \{1, ..., J\}$  and  $\ell \in \{1, ..., s\}$ such that

$$(M,R) = \left(\frac{N-\ell+1}{s}, \frac{s-1}{2} + \frac{\ell(\ell-1)}{2s}\right).$$
 (54)

Based on the parameter values, we prove  $R_{dec}(M) \leq$ 2.00884R by considering the following 3 possible scenarios: a). If  $N \ge 9s$ , we first have the follows given (20):

$$R_{\rm dec}(M) \le \frac{N-M}{M}.$$
(55)

Due to (54), the above inequality is equivalent to

$$R_{\rm dec}(M) \le s - 1 + \frac{s(\ell - 1)}{N - \ell + 1}.$$
(56)

Recall that  $s \ge \ell$  and  $N \ge 9s$ , we have

$$R_{\rm dec}(M) \le s - 1 + \frac{s(\ell - 1)}{N - s} \le s - 1 + \frac{\ell - 1}{8}.$$
(57)

$$R_{\rm dec}(M) \le s - 1 + \frac{\sqrt{\ell} - 1}{8} \cdot \sqrt{\frac{(s-1)}{s}} \cdot \ell \\ = s - 1 + 2 \cdot \sqrt{\frac{s-1}{256}} \cdot \sqrt{\frac{\ell(\ell-1)}{s}}.$$
(58)

Applying the AM-GM inequality to the second term of the RHS, we have

$$R_{\rm dec}(M) \le s - 1 + \frac{s - 1}{256} + \frac{\ell(\ell - 1)}{s}.$$
 (59)

Because  $\ell \geq 1$ , we can thus upper bound  $R_{dec}(M)$  as a function of R, which is given in (54):

$$R_{\text{dec}}(M) \le (2 + \frac{1}{128})(\frac{s-1}{2} + \frac{\ell(\ell-1)}{2s}) \le 2.00884R.$$
(60)

b). If N < 9s and  $N \leq 81$ , we upper bound  $R_{dec}(M)$  as follows:

$$R_{\rm dec}(M) \le \frac{N-M}{M} (1 - (1 - \frac{M}{N})^N).$$
 (61)

Note that both the above bound and R are functions of N, sand  $\ell$ , which can only take values from  $\{1, ..., 81\}$ . Through a brute-force search, we can show that  $R_{\rm dec}(M) \leq 2.000 R \leq$ 2.00884R.

c). If N < 9s and N > 81, recall that  $M = \frac{N - \ell + 1}{s}$  from (54), we have

$$M \le \frac{N}{s} < 9. \tag{62}$$

$$R = \frac{s-1}{2} + \frac{(N-sM)(N-sM+1)}{2s}$$
$$= \frac{(1+M^2)s}{2} + \frac{N(N+1)}{2s} - (N+\frac{1}{2})M - \frac{1}{2}.$$
 (63)

Applying the AM-GM inequality to the first two terms of the RHS, we have

$$R \ge \sqrt{(1+M^2)N(N+1)} - (N+\frac{1}{2})M - \frac{1}{2}.$$
 (64)

From (62),  $N > 81 > M^2$ , we have  $\sqrt{N(N+1)} \ge \sqrt{M^2(M^2+1)} + N - M^2$ . Consequently,

$$R \ge \sqrt{1 + M^2} (\sqrt{M^2(M^2 + 1)} + N - M^2) - (N + \frac{1}{2})M - \frac{1}{2} = (N - 81)(\sqrt{1 + M^2} - M) + (81 - M^2)(\sqrt{1 + M^2} - M) + \frac{M - 1}{2}.$$
 (65)

On the other hand, we upper bound  $R_{dec}(M)$  as follows:

$$R_{\rm dec}(M) \le \frac{N-M}{M} (1 - (1 - \frac{M}{N})^N) = \frac{N-M}{M} (1 - e^{\ln(1 - \frac{M}{N})N}).$$
(66)

From (62),  $\frac{M}{N} < \frac{9}{81} = \frac{1}{9}$ , it is easy to show that  $\ln(1 - \frac{M}{N}) \ge -\frac{M}{N} - \frac{9}{16} \left(\frac{M}{N}\right)^2$ . Hence,

$$R_{\text{dec}}(M) \leq \frac{N-M}{M} \left(1 - e^{-M - \frac{9}{16}\frac{M^2}{N}}\right)$$
$$\leq \frac{N-M}{M} \left(1 - e^{-M} \left(1 - \frac{9}{16}\frac{M^2}{N}\right)\right)$$
$$\leq \frac{N-M}{M} \left(1 - e^{-M}\right) + \frac{N}{M} e^{-M}\frac{9}{16}\frac{M^2}{N}$$
$$= (N-81)\frac{1 - e^{-M}}{M} + \frac{81 - M}{M} \left(1 - e^{-M}\right)$$
$$+ \frac{9}{16}Me^{-M}. \tag{67}$$

Numerically, we can verify that the following inequalities hold for  $M \in [0, 9)$ :

$$\frac{1-e^{-M}}{M} \le 2.00884(\sqrt{1+M^2}-M), \tag{68}$$
$$\frac{81-M}{M}(1-e^{-M}) + \frac{9}{16}Me^{-M} \le 2.00884\left((81-M^2)(\sqrt{1+M^2}-M) + \frac{M-1}{2}\right). \tag{69}$$

Hence when N > 81, by computing  $(N - 81) \times (68) + (69)$ , we have  $R_{\text{dec}}(M) \leq 2.00884R$ .

To conclude,  $R_{dec}(M) \leq 2.00884R$  holds for any  $(M, R) \in S_{Lower}$  for all three cases. This completes the proof of Lemma 1.

## Appendix D Proof of Lemma 2

If R is  $\epsilon$ -achievable, we can find message  $X_d$  such that for each user k,  $W_{d_k}$  can be decoded from  $Z_k$  and  $X_d$  with probability of error of at most  $\epsilon$ . Using Fano's inequality, the following bound holds:

$$H(W_{d_k}|Z_k, X_d) \le 1 + \epsilon F \qquad \forall k \in \{1, \dots, K\}.$$
(70)

Equivalently,

$$H(X_{d}|Z_{k}) \ge H(W_{d_{k}}|Z_{k}) + H(X_{d}|W_{d_{k}}, Z_{k}) - (1 + \epsilon F) \quad \forall k \in \{1, ..., K\}.$$
(71)

Note that the LHS of the above inequality lower bounds the communication load. If we lower bound the term  $H(X_d|W_{d_k}, Z_k)$  on the RHS by 0, we obtain the single user cutset bound. However, we enhance this cutset bound by bounding  $H(X_d|W_{d_k}, Z_k)$  with non-negative functions. On a high level, we view  $H(X_d|W_{d_k}, Z_k)$  as the communication load on an enhanced caching system, where  $W_{d_k}$  and  $Z_k$  are known by all the users. Using similar approach, we can lower bound  $H(X_d|W_{d_k}, Z_k)$  by the sum of a single cutset bound on this enhanced system, and another entropy function that can be interpreted as the communication load on a further enhanced system. We can recursively apply this bounding technique until all user demands are publicly known.

$$H(W_{d_k}|Z_{\{1,...,k\}}, X_d, W_{\{d_1,...,d_{k-1}\}}) \le 1 + \epsilon F$$
  
$$\forall k \in \{1,...,K\}.$$
(72)

Equivalently,

$$H(X_{d}|Z_{\{1,...,k\}}, W_{\{d_{1},...,d_{k-1}\}}) \geq H(W_{d_{k}}|Z_{\{1,...,k\}}, W_{\{d_{1},...,d_{k-1}\}}) + H(X_{d}|Z_{\{1,...,k\}}, W_{\{d_{1},...,d_{k}\}}) - (1 + \epsilon F) \quad \forall k \in \{1,...,K\}.$$
(73)

Adding the above inequality for  $k \in \{1, ..., \min\{N, K\}\}$ , we have

$$H(X_{d}|Z_{\{1\}}) \geq \min\{N,K\} \\ \sum_{k=1}^{\min\{N,K\}} \left( H(W_{d_{k}}|Z_{\{1,\dots,k\}}, W_{\{d_{1},\dots,d_{k-1}\}}) - (1+\epsilon F) \right) \\ + H(X_{d}|Z_{\{1,\dots,\min\{N,K\}\}}, W_{\{d_{1},\dots,d_{\min\{N,K\}}\}}) \\ \geq \sum_{k=1}^{\min\{N,K\}} H(W_{d_{k}}|Z_{\{1,\dots,k\}}, W_{\{d_{1},\dots,d_{k-1}\}}) \\ - \min\{N,K\}(1+\epsilon F).$$
(74)

Thus, R is bounded by

$$R \geq \frac{1}{F} H(X_{d}|Z_{\{1\}})$$

$$\geq \frac{1}{F} \left(\sum_{k=1}^{\min\{N,K\}} H(W_{d_{k}}|Z_{\{1,\dots,k\}}, W_{\{d_{1},\dots,d_{k-1}\}})\right)$$

$$-\min\{N,K\}(\frac{1}{F}+\epsilon).$$
(75)

One can show that this approach strictly improves the compound cutset bound, which was used in most of the prior works.

#### APPENDIX E Proof of Lemma 3

In this appendix, we prove that for any prefetching scheme  $\phi$ , the rate  $R_A(F, \phi)$  is lower bounded by the RHS of (33), for any parameters s and  $\alpha$ . Now we consider any such  $s \in$ 

 $\{1, ..., \min\{N, K\}\}$  and  $\alpha \in [0, 1]$ . From the definition of  $R_{\rm A}(F, \phi)$  and the non-negativity of entropy functions, we have

$$R_{\rm A}^*(F,\phi)F \ge \left(\sum_{k=1}^{s-1} H^*(W_k|Z_{\{1,\dots,k\}}, W_{\{1,\dots,k-1\}})\right) + \alpha H^*(W_s|Z_{\{1,\dots,s\}}, W_{\{1,\dots,s-1\}}).$$
(76)

Each term in the above lower bound can be bounded in the following 2 ways: <sup>12</sup>

$$H^{*}(W_{k}|Z_{\{1,...,k\}}, W_{\{1,...,k-1\}}) \geq \frac{H^{*}(W_{\{k,...,N\}}|Z_{\{1,...,k\}}, W_{\{1,...,k-1\}})}{N-k+1} \geq F - \frac{H^{*}(Z_{\{1,...,k\}}|W_{\{1,...,k-1\}})}{N-k+1}$$
(77)

$$H^{*}(W_{k}|Z_{\{1,...,k\}}, W_{\{1,...,k-1\}}) = F - H^{*}(Z_{\{1,...,k\}}|W_{\{1,...,k-1\}}) + H^{*}(Z_{\{1,...,k\}}|W_{\{1,...,k\}}) \\ \geq F - H^{*}(Z_{\{1,...,k\}}|W_{\{1,...,k-1\}}) + \frac{k}{k+1}H^{*}(Z_{\{1,...,k+1\}}|W_{\{1,...,k\}})$$
(78)

We aim to use linear combinations of the above two bounds in (76), such that the coefficient of each  $H^*(Z_{\{1,\ldots,k\}}|W_{\{1,\ldots,k-1\}})$  in the resulting lower bound is 0 for all but one k. To do so, we construct the following sequences:

$$a_x = \frac{2\alpha s + s(s-1) - (x+1)x}{2x(N-x)},$$
(79)

$$b_x = \frac{2\alpha s + s(s-1) - x(x-1)}{2x(N-x+1)}.$$
(80)

We can verify that these sequences satisfy the following equations:

$$\frac{1-a_x}{N-x+1} + a_x = b_x,$$
(81)

$$\frac{x}{x+1}a_x = b_{x+1}.$$
 (82)

Let  $\ell \in \{1, ..., s\}$  be the minimum value such that (12) holds, we can prove that  $a_x \in [0, 1]$  for  $x \in \{\ell, ..., s-1\}$ . Because  $\ell$  is the minimum of such values, we can also prove that  $b_l \geq \frac{\ell-1}{\ell}$ . Using the above properties of sequences a and b, we lower bound  $R_{\rm A}(F, \phi)$  as follows:

For each  $x \in \{\ell, ..., s-1\}$ , by computing  $(1-a_x) \times (77) +$  $a_x \times (78)$ , we have

$$\begin{split} H^*(W_x|Z_{\{1,...,x\}},W_{\{1,...,x-1\}}) \\ &\geq (1-a_x)\left(F-\frac{H^*(Z_{\{1,...,k\}}|W_{\{1,...,k-1\}})}{N-k+1}\right) \\ &\quad +a_x(F-H^*(Z_{\{1,...,k\}}|W_{\{1,...,k-1\}}) \\ &\quad +\frac{k}{k+1}H^*(Z_{\{1,...,k+1\}}|W_{\{1,...,k\}})) \\ &= F-(\frac{1-a_x}{N-x+1}+a_x)H^*(Z_{\{1,...,x\}}|W_{\{1,...,x-1\}}) \\ &\quad +a_x\frac{x}{x+1}H^*(Z_{\{1,...,x+1\}}|W_{\{1,...,x\}}) \end{split}$$

<sup>12</sup>Rigorously, (78) requires k < K. However, we will only apply this bound for k < s, which satisfies this condition.

$$=F - b_x H^*(Z_{\{1,...,x\}} | W_{\{1,...,x-1\}}) + b_{x+1} H^*(Z_{\{1,...,x+1\}} | W_{\{1,...,x\}}).$$
(83)

Moreover, we have the follows from (77):

$$\alpha H^{*}(W_{s}|Z_{\{1,\ldots,s\}}, W_{\{1,\ldots,s-1\}}) \\ \geq \alpha \left(F - \frac{H^{*}(Z_{\{1,\ldots,s\}}|W_{\{1,\ldots,s-1\}})}{N-s+1}\right) \\ = \alpha F - b_{s}H^{*}(Z_{\{1,\ldots,s\}}|W_{\{1,\ldots,s-1\}}).$$
(84)

Consequently,

 $\mathbf{k}$ 

 $R_{\rm A}$ 

$$\sum_{k=\ell}^{s-1} H^*(W_k | Z_{\{1,\dots,k\}}, W_{\{1,\dots,k-1\}}) + \alpha H^*(W_s | Z_{\{1,\dots,s\}}, W_{\{1,\dots,s-1\}}) \ge (s-\ell+\alpha)F - b_\ell H^*(Z_{\{1,\dots,\ell\}} | W_{\{1,\dots,\ell-1\}}).$$
(85)

On the other hand,

$$\sum_{k=1}^{\ell-1} H^*(W_k | Z_{\{1,\dots,k\}}, W_{\{1,\dots,k-1\}}) \\ \ge \sum_{k=1}^{\ell-1} (F - H^*(Z_{\{1,\dots,k\}} | W_{\{1,\dots,k-1\}})) \\ + \frac{k}{k+1} H^*(Z_{\{1,\dots,k+1\}} | W_{\{1,\dots,k\}})) \\ = \sum_{k=1}^{\ell-1} (F - \frac{1}{k} H^*(Z_{\{1,\dots,\ell\}} | W_{\{1,\dots,\ell-1\}})) \\ + \frac{\ell-1}{\ell} H^*(Z_{\{1,\dots,\ell\}} | W_{\{1,\dots,\ell-1\}})) \\ \ge (\ell-1)F - (\ell-1)MF \\ + \frac{\ell-1}{\ell} H^*(Z_{\{1,\dots,\ell\}} | W_{\{1,\dots,\ell-1\}}).$$
(86)

Combining (76), (85), and (86), we have

$$(F, \phi)F \ge (\ell - 1)F - (\ell - 1)MF + \frac{\ell - 1}{\ell}H^*(Z_{\{1,...,\ell\}}|W_{\{1,...,\ell-1\}}) + (s - \ell + \alpha)F - b_{\ell}H^*(Z_{\{1,...,\ell\}}|W_{\{1,...,\ell-1\}}) = (s - 1 + \alpha)F - (\ell - 1)MF + \left(\frac{\ell - 1}{\ell} - b_l\right)H^*(Z_{\{1,...,\ell\}}|W_{\{1,...,\ell-1\}}).$$
(87)

Recall that  $b_{\ell} \geq \frac{\ell-1}{\ell}$ , we have

$$R_{A}(F,\phi)F \ge (s-1+\alpha)F - (\ell-1)MF - (b_{\ell} - \frac{\ell-1}{\ell})\ell MF = (s-1+\alpha)F - \frac{s(s-1) - \ell(\ell-1) + 2\alpha s}{2(N-\ell+1)}MF.$$
(88)

This completes the proof of Lemma 3.

#### APPENDIX F **PROOF OF LEMMA 5**

To simplify the discussion, we adopt the notation of  $H^*(W_{\mathcal{A}}, Z_{\mathcal{B}})$  which is defined in the proof of Theorem 2. Moreover, we generalize this notation to include the variables for the messages  $X_d$ . For any permutations  $p \in \mathcal{P}_N$ ,  $q \in \mathcal{P}_K$ 

and for any demand  $\boldsymbol{d} \in \{1, ..., N\}^K$ , we define  $\boldsymbol{d}(p,q)$  be a demand where for each  $k \in \{1, ..., K\}$ , user q(k) requests file  $p(d_k)$ . Then for any subset for demands  $\mathcal{D} \subseteq \{1, ..., N\}^K$ , we define  $\mathcal{D}(p,q) = \{\boldsymbol{d}(p,q) | \boldsymbol{d} \in \mathcal{D}\}$ . Now for any subsets  $\mathcal{A} \subseteq \{1, ..., N\}, \mathcal{B} \subseteq \{1, ..., K\}$ , and  $\mathcal{D} \subseteq \{1, ..., N\}^K$ , we define

$$H^{*}(X_{\mathcal{D}}, W_{\mathcal{A}}, Z_{\mathcal{B}}) \triangleq \frac{1}{N!K!} \sum_{p \in \mathcal{P}_{N}, q \in \mathcal{P}_{K}} H(X_{\mathcal{D}(p,q)}, W_{p\mathcal{A}}, Z_{q\mathcal{B}}).$$
(89)

For any  $i \in \{1, ..., n\}$  and  $j \in \{1, ..., \alpha\}$  let  $d^{i,j}$  be a demand satisfying

$$d_{l}^{i,j} = \begin{cases} l - i + (j-1)(K-n) + \beta \\ & \text{if } i+1 \le l \le i+K-n, \\ 1 & \text{otherwise.} \end{cases}$$
(90)

Note that for all demands  $d^{i,j}$ , user 1 requests file 1, hence we have

$$H(W_1|X_{\boldsymbol{d}^{i,j}}, Z_1) \le 1 + \epsilon F \tag{91}$$

using Fano's inequality. Consequently,

$$RF \ge H(X_{d^{i,j}})$$
  

$$\ge H(X_{d^{i,j}}|Z_1) + H(W_1|X_{d^{i,j}}, Z_1) - (1 + \epsilon F)$$
  

$$= H(W_1|Z_1) + H(X_{d^{i,j}}|W_1, Z_1) - (1 + \epsilon F). \quad (92)$$

Due to the homogeneity of the problem, we have

 $RF \ge H^*(W_1|Z_1) + H^*(X_{d^{i,j}}|W_1, Z_1) - (1 + \epsilon F).$  (93) For each  $i \in \{1, ..., n\}, j \in \{1, ..., \alpha\}$ , and  $k \in \{1, ..., i\}$ , we have the following identity:

$$H^*(X_{\mathbf{d}^{i,j}}|W_1, Z_1) = H^*(X_{\mathbf{d}^{i,j}}|W_1, Z_k).$$
(94)

Hence, we have

$$RF \ge H^*(W_1|Z_1) + \frac{2}{n(n+1)\alpha} \sum_{k=1}^n \sum_{i=k}^n \sum_{j=1}^\alpha H^*(X_{d^{i,j}}|W_1, Z_k) - (1 + \epsilon F).$$
(95)

For  $k \in \{1, ..., n\}$ , let  $\mathcal{D}_k$  and  $\mathcal{D}_k^+$  denote the following set of demands:

$$\mathcal{D}_k = \{ \boldsymbol{d}^{k,j} | j \in \{1, \dots, \alpha\} \},$$
(96)

$$\mathcal{D}_k^+ = \bigcup_{i=k} \mathcal{D}_i,\tag{97}$$

we have

$$RF \ge H^*(W_1|Z_1) + \frac{2}{n(n+1)\alpha} \sum_{k=1}^n H^*(X_{\mathcal{D}_k^+}|W_1, Z_k) - (1+\epsilon F) \ge H^*(W_1|Z_1) + \frac{2}{n(n+1)\alpha} \sum_{k=1}^n H^*(X_{\mathcal{D}_k^+}|W_{\{1,\dots,\beta\}}, Z_k) - (1+\epsilon F) \ge H^*(W_1|Z_1) + \frac{2}{n(n+1)\alpha} \sum_{k=1}^n \left( H^*(Z_k, X_{\mathcal{D}_k^+}|W_{\{1,\dots,\beta\}}) \right)$$

$$-H^{*}(Z_{k}|W_{\{1,\ldots,\beta\}})) - (1+\epsilon F).$$
(98)

To further bound R, we only need a lower bound for  $\sum_{k=1}^{n} H^*(Z_k, X_{\mathcal{D}_k^+} | W_{\{1,...,\beta\}})$ , which is derived as follows:

For each  $i \in \{1, ..., K - n\}$ , let  $S_i$  be subset of files defined as follows:

$$S_i = \{i + (j-1)(K-n) + \beta \mid j \in \{1, ..., \alpha\}\}.$$
 (99)

From the decodability constraint, for any  $k \in \{1, ..., n\}$ , each file in  $S_i$  can be decoded by user i + k given  $X_{\mathcal{D}_k}$ . Using Fano's inequality, we have

$$H^*(W_{\mathcal{S}_i}|X_{\mathcal{D}_k}, Z_{i+k}) \le \alpha(1+\epsilon F).$$
(100)

Let  $\mathcal{S}_i^-$  be subset of files defined as follows

$$\mathcal{S}_i^- = \left(\bigcup_{j=1}^i \mathcal{S}_j\right) \bigcup \{1, ..., \beta\}.$$
 (101)

We have

$$0 \geq H^{*}(W_{\mathcal{S}_{i}}|X_{\mathcal{D}_{k}^{+}}, Z_{i+k}, W_{\mathcal{S}_{i-1}^{-}}) - \alpha(1 + \epsilon F)$$
  
=  $H^{*}(X_{\mathcal{D}_{k}^{+}}, Z_{i+k}|W_{\mathcal{S}_{i}}, W_{\mathcal{S}_{i-1}^{-}}) + H^{*}(W_{\mathcal{S}_{i}}|W_{\mathcal{S}_{i-1}^{-}})$   
 $- H^{*}(X_{\mathcal{D}_{k}^{+}}, Z_{i+k}|W_{\mathcal{S}_{i-1}^{-}}) - \alpha(1 + \epsilon F)$   
=  $H^{*}(X_{\mathcal{D}_{k}^{+}}, Z_{i+k}|W_{\mathcal{S}_{i}^{-}}) + \alpha F - H^{*}(X_{\mathcal{D}_{k}^{+}}, Z_{i+k}|W_{\mathcal{S}_{i-1}^{-}})$   
 $- \alpha(1 + \epsilon F).$  (102)

Consequently,

$$0 \geq \sum_{k=1}^{n} \sum_{i=1}^{K-n} \left( H^{*}(X_{\mathcal{D}_{k}^{+}}, Z_{i+k} | W_{\mathcal{S}_{i}^{-}}) + \alpha F - H^{*}(X_{\mathcal{D}_{k}^{+}}, Z_{i+k} | W_{\mathcal{S}_{i-1}^{-}}) - \alpha(1 + \epsilon F) \right)$$

$$= \sum_{k=1}^{n} \left( \sum_{i=1}^{K-n} \left( H^{*}(X_{\mathcal{D}_{k}^{+}}, Z_{i+k-1} | W_{\mathcal{S}_{i-1}^{-}}) - H^{*}(X_{\mathcal{D}_{k}^{+}}, Z_{i+k} | W_{\mathcal{S}_{i-1}^{-}}) \right) + H^{*}(X_{\mathcal{D}_{k}^{+}}, Z_{K-n+k} | W_{\mathcal{S}_{n}^{-}}) - H^{*}(X_{\mathcal{D}_{k}^{+}}, Z_{k} | W_{\mathcal{S}_{0}^{-}}) \right)$$

$$+ \alpha n(K - n)(F - 1 - \epsilon F)$$

$$\geq \sum_{k=1}^{n} \left( \sum_{i=1}^{K-n} \left( H^{*}(X_{\mathcal{D}_{k}^{+}}, Z_{i+k-1} | W_{\mathcal{S}_{i-1}^{-}}) - H^{*}(X_{\mathcal{D}_{k}^{+}}, Z_{i+k} | W_{\mathcal{S}_{i-1}^{-}}) \right) - H^{*}(X_{\mathcal{D}_{k}^{+}}, Z_{i+k} | W_{\mathcal{S}_{i-1}^{-}}) \right)$$

$$+ \alpha n(K - n)(F - 1 - \epsilon F). \tag{103}$$

Hence, we obtain the following lower bound:

$$\sum_{k=1}^{N} H^*(X_{\mathcal{D}_k^+}, Z_k | W_{\mathcal{S}_0^-}))$$

$$\geq \sum_{k=1}^{n} \sum_{i=1}^{K-n} \left( H^*(X_{\mathcal{D}_k^+}, Z_{i+k-1} | W_{\mathcal{S}_{i-1}^-}) - H^*(X_{\mathcal{D}_k^+}, Z_{i+k} | W_{\mathcal{S}_{i-1}^-}) \right)$$

$$+ \alpha n(K-n)(F-1-\epsilon F)$$

$$= \sum_{i=1}^{K-n} \sum_{k=1}^{n} \left( H^*(X_{\mathcal{D}_k^+}, Z_{i+k-1} | W_{\mathcal{S}_{i-1}^-}) - H^*(X_{\mathcal{D}_k^+}, Z_{i+k} | W_{\mathcal{S}_{i-1}^-}) \right) + \alpha n(K-n)(F-1-\epsilon F)$$

$$= \sum_{i=1}^{K-n} \sum_{k=1}^{n} \left( H^*(Z_{i+k-1} | X_{\mathcal{D}_k^+}, W_{\mathcal{S}_{i-1}^-}) - H^*(Z_{i+k} | X_{\mathcal{D}_k^+}, W_{\mathcal{S}_{i-1}^-}) \right) + \alpha n(K-n)(F-1-\epsilon F)$$
(104)

Note that  $\mathcal{D}_k^+ \subseteq \mathcal{D}_{k-1}^+$ , we have  $H^*(Z_{i+k}|X_{\mathcal{D}_{k+1}^+}, W_{\mathcal{S}_{i-1}^-}) \ge H^*(Z_{i+k}|X_{\mathcal{D}_k^+}, W_{\mathcal{S}_{i-1}^-})$ . Consequently,

$$\sum_{k=1}^{n} H^{*}(X_{\mathcal{D}_{k}^{+}}, Z_{k} | W_{\mathcal{S}_{0}^{-}}))$$

$$\geq \sum_{i=1}^{K-n} \left( H^{*}(Z_{i} | X_{\mathcal{D}_{1}^{+}}, W_{\mathcal{S}_{i-1}^{-}}) - H^{*}(Z_{i+n} | X_{\mathcal{D}_{n}^{+}}, W_{\mathcal{S}_{i-1}^{-}}) \right)$$

$$+ \alpha n(K - n)(F - 1 - \epsilon F)$$

$$\geq -\sum_{i=1}^{K-n} H^{*}(Z_{i+n} | W_{\mathcal{S}_{i-1}^{-}}) + \alpha n(K - n)(F - 1 - \epsilon F)$$

$$= -\sum_{i=1}^{K-n} H^{*}(Z_{1} | W_{\{1, \dots, \beta + i\alpha\}})$$

$$+ \alpha n(K - n)(F - 1 - \epsilon F). \tag{105}$$

Applying (105) to (98), we have

$$RF \ge H^{*}(W_{1}|Z_{1}) + \frac{2}{n(n+1)\alpha} \left( \alpha n(K-n)(F-1-\epsilon F) - \sum_{k=1}^{n} H^{*}(Z_{k}|W_{\{1,...,\beta\}}) - \sum_{k=0}^{K-n-1} H^{*}(Z_{1}|W_{\{1,...,\beta+i\alpha\}}) \right) - (1+\epsilon F)$$

$$=H^{*}(W_{1}|Z_{1}) + \frac{2}{n(n+1)\alpha} \left( \alpha n(K-n)F - nH^{*}(Z_{1}|W_{\{1,...,\beta\}}) - \sum_{i=0}^{K-n-1} H^{*}(Z_{1}|W_{\{1,...,\beta+i\alpha\}}) \right) - \frac{2K-n+1}{n+1} (1+\epsilon F).$$
(106)

#### APPENDIX G PROOF OF THEOREM 1 FOR AVERAGE RATE

Here we prove Theorem 1 for the average rate (i.e. inequalities (8) and (10)). The upper bounds of  $R_{\text{ave}}^*$  in these inequalities can be achieved using the caching scheme provided in [24], hence we only need to prove their lower bounds. To do so, we define the following terminology:

We divide the set of all demands, denoted by  $\mathcal{D}$ , into smaller subsets, and refer them to as *types*. We use the same definition

in [24], which are stated as follows: Given an arbitrary demand d, we define its *statistics*, denoted by s(d), as a sorted array of length N, such that  $s_i(d)$  equals the number of users that request the *i*th most requested file. We denote the set of all possible statistics by S. Grouping by the same statistics, the set of all demands D can be broken into many subsets. For any statistics  $s \in S$ , we define type  $D_s$  as the set of queries with statistics s. Note that for each demand d, the value  $N_e(d)$  only depends on its statistics s(d), and thus the value is identical across all demands in  $D_s$ . For convenience, we denote that value by  $N_e(s)$ .

Given a prefetching scheme  $\phi$  and a type  $\mathcal{D}_s$ , we say a rate R is  $\epsilon$ -achievable for type  $\mathcal{D}_s$  if we can find a function R(d) that is  $\epsilon$ -achievable for any demand d in  $\mathcal{D}_s$ , satisfying  $R = \mathbb{E}_d[R(d)]$ , where d is uniformly random in  $\mathcal{D}_s$ . Hence, to characterize  $R_{\text{ave}}^*$ , it is sufficient to lower bound the  $\epsilon$ achievable rates for each type individually, and show that for each type, the caching scheme provided in [24] is within the given constant factors optimal for large F and small  $\epsilon$ .

We first lower bound any  $\epsilon$ -achievable rate for each type as follows : Within a type  $\mathcal{D}_s$ , we can find a demand d, such that users in  $\{1, ..., N_e(s)\}$  requests different files. We can easily generalize Lemma 2 to this demand, and any  $\epsilon$  achievable rate of this demand, denoted by  $R_d$ , is lower bounded by the following inequality:

$$R_{d} \geq \frac{1}{F} \left( \sum_{k=1}^{N_{e}(s)} H(W_{d_{k}} | Z_{\{1,...,k\}}, W_{\{d_{1},...,d_{k-1}\}}) \right) - N_{e}(s)(\frac{1}{F} + \epsilon).$$
(107)

Applying the same bounding technique to all demands in type  $\mathcal{D}_s$ . We can prove that any rate that is  $\epsilon$ -achievable for  $\mathcal{D}_s$ , denoted by  $R_s$ , is bounded by the follows:

$$R_{s} \geq \frac{1}{F} \left( \sum_{k=1}^{N_{e}(s)} H^{*}(W_{k} | Z_{\{1,...,k\}}, W_{\{1,...,k-1\}}) \right) - N_{e}(s)(\frac{1}{F} + \epsilon),$$
(108)

where function  $H^*(\cdot)$  is defined in the proof of Theorem 2.

Following the same steps in the proof of Theorem 2, we can prove that

$$R_{s} \ge s - 1 + \alpha - \frac{s(s-1) - \ell(\ell-1) + 2\alpha s}{2(N-\ell+1)} M - N_{e}(s)(\frac{1}{F} + \epsilon),$$
(109)

for arbitrary  $s \in \{1, ..., N_e(s)\}$ ,  $\alpha \in [0, 1]$ , where  $\ell \in \{1, ..., s\}$  is the minimum value such that

$$\frac{s(s-1) - \ell(\ell-1)}{2} + \alpha s \le (N - \ell + 1)\ell.$$
(110)

On the other hand, the caching scheme provided in [24] achieves an average rate of  $\operatorname{Conv}\left(\frac{\binom{K}{r+1} - \binom{K-N_{\rm e}(s)}{r+1}}{\binom{K}{r}}\right)$  within each type  $\mathcal{D}_s$ . Using the results in [24], we can easily prove that this average rate can be upper bounded by  $R_{\rm dec}(M, s)$ , defined as

$$R_{\rm dec}(M, s) \triangleq \frac{N - M}{M} (1 - (1 - \frac{M}{N})^{N_{\rm e}(s)}).$$
(111)

Hence, in order to prove (8) and (10), it suffices to prove that for large F and small  $\epsilon$ , any  $\epsilon$ -achievable rate  $R_s$  for any type

 $\mathcal{D}_s$  satisfies  $R_s \geq R_{\text{dec}}(M, s)/2.00884$  in the general case, and  $R_s \geq R_{\text{dec}}(M, s)/2$  when  $N \geq \frac{K(K+1)}{2}$ .

Note that the above characterization of  $R_s$  exactly matches a characterization of  $R^*$  for a caching system with N files and  $N_{\rm e}(s)$  users. Specifically, the lower bound of  $R_s$  given by (109) exactly matches Theorem 2, and the upper bound  $R_{\rm dec}(M, s)$  defined in (111) exactly matches the upper bound  $R_{\rm dec}(M)$  defined in (20). Thus, by reusing the same arguments in the proof of Theorem 1 for the peak rate, we can easily prove that  $R_s \ge R_{\rm dec}(M, s)/2.00884$  holds for the general case, and  $R_s \ge R_{\rm dec}(M, s)/2$  holds for sufficiently large Nwhen  $\frac{N_{\rm e}(s)M}{N} > 1$ . Hence, to prove Theorem 1 for the average rate, we only need  $R_s \ge R_{\rm dec}(M, s)/2$  for sufficiently large N to also hold when  $\frac{N_{\rm e}(s)M}{N} \le 1$ , which can be easily proved as follows:

Using the same arguments in the proof of Theorem 1 for the peak rate, the following inequality can be derived from (109) for large N, large F and small  $\epsilon$ :

$$R_{\boldsymbol{s}} \ge N_{\boldsymbol{e}}(\boldsymbol{s}) - \frac{N_{\boldsymbol{e}}(\boldsymbol{s})(N_{\boldsymbol{e}}(\boldsymbol{s})+1)}{2} \cdot \frac{M}{N}, \qquad (112)$$

which is a linear function of M. Furthermore, since  $R_{dec}(M, s)$  is convex, we only need to check that

$$\frac{R_{\text{dec}}(M, \boldsymbol{s})}{2} \le N_{\text{e}}(\boldsymbol{s}) - \frac{N_{\text{e}}(\boldsymbol{s})(N_{\text{e}}(\boldsymbol{s})+1)}{2} \cdot \frac{M}{N}$$
(113)

holds at  $\frac{N_{e}(s)M}{N} \in \{0, 1\}$ . For  $\frac{N_{e}(s)M}{N} = 0$ , we have

$$\frac{1}{\frac{N}{2}} = 0, \text{ we have}$$

$$\frac{R_{\text{dec}}(M, \boldsymbol{s})}{2} = \frac{N_{\text{e}}(\boldsymbol{s})}{2} \le N_{\text{e}}(\boldsymbol{s}) =$$

$$N_{\text{e}}(\boldsymbol{s}) - \frac{N_{\text{e}}(\boldsymbol{s})(N_{\text{e}}(\boldsymbol{s})+1)}{2} \cdot \frac{M}{N}. \quad (114)$$

For  $\frac{N_{e}(s)M}{N} = 1$ , we have

$$\frac{R_{\text{dec}}(M, \boldsymbol{s})}{2} = \frac{N_{\text{e}}(\boldsymbol{s}) - 1}{2} \left( 1 - \left( 1 - \frac{1}{N_{\text{e}}(\boldsymbol{s})} \right)^{N_{\text{e}}(\boldsymbol{s})} \right)$$
$$\leq \frac{N_{\text{e}}(\boldsymbol{s}) - 1}{2}$$
$$= N_{\text{e}}(\boldsymbol{s}) - \frac{N_{\text{e}}(\boldsymbol{s})(N_{\text{e}}(\boldsymbol{s}) + 1)}{2} \cdot \frac{M}{N}. \quad (115)$$

This completes the proof of Theorem 1.

## APPENDIX H THE EXACT RATE-MEMORY TRADEOFF FOR TWO-USER

CASE

As mentioned in Remark 4, we can completely characterize the rate-memory tradeoff for average rate for the two-user case, for any possible values of N and M. We formally state this result in the following corollary:

**Corollary 1.** For a caching system with 2 users, a database of N files, and a local cache size of M files at each user, we have

$$R_{\text{ave}}^* = R_{\text{u,ave}}(N, K, r), \qquad (116)$$

where  $R_{u,ave}(N, K, r)$  is defined in Definition 1.

*Proof.* For the single-file case, only one possible demand exists. The average rate thus equals the peak rate, which can be easily characterized. Hence, we omit the proof and focus

on cases where  $N \geq 2$ . Note that  $R_{u,ave}$  can be achieved using the scheme provided in [24], we only need to prove that  $R_{ave}^* \geq R_{u,ave}(N, K, r)$ .

As shown in Appendix G, the average rate within each type  $\mathcal{D}_s$  is bounded by (108). Hence, the minimum average rate under uniform file popularity given a prefetching scheme  $\phi$ , denoted by  $R(\phi)$ , is lower bounded by

$$R(\phi) \ge \mathbb{E}_{s} \left[ \frac{1}{F} \left( \sum_{k=1}^{N_{c}(s)} H^{*}(W_{k} | Z_{\{1,...,k\}}, W_{\{1,...,k-1\}}) \right) - N_{e}(s)(\frac{1}{F} + \epsilon) \right].$$
(117)

Note that for the two-user case,  $N_e(s)$  equals 1 with probability  $\frac{1}{N}$ , and 2 with probability  $\frac{N-1}{N}$ . Consequently,

$$R(\phi) \ge \frac{1}{F} \left( H^*(W_1|Z_1) + \frac{N-1}{N} \cdot H^*(W_2|Z_{\{1,2\}}, W_1) \right) - \frac{2N-1}{N} \cdot (\frac{1}{F} + \epsilon).$$
(118)

Using the technique developed in proof of Theorem 2, we have the following two lower bounds

$$R(\phi) \geq \frac{1}{F} H^{*}(W_{1}|Z_{1}) - \frac{2N-1}{N} \cdot (\frac{1}{F} + \epsilon)$$

$$\geq 1 - \frac{M}{N} - \frac{2N-1}{N} \cdot (\frac{1}{F} + \epsilon), \qquad (119)$$

$$R(\phi) \geq \frac{1}{F} \left( H^{*}(W_{1}|Z_{1}) + \frac{N-1}{N} \cdot H^{*}(W_{2}|Z_{\{1,2\}}, W_{1}) \right)$$

$$- \frac{2N-1}{N} \cdot (\frac{1}{F} + \epsilon)$$

$$\geq \frac{1}{F} \left( H^{*}(W_{1}|Z_{1}) + \frac{1}{N} \cdot ((N-1)F - 2H^{*}(Z_{1}|W_{1})) \right)$$

$$- \frac{2N-1}{N} \cdot (\frac{1}{F} + \epsilon)$$

$$\geq \frac{2N-1}{N} - \frac{3N-2}{N} \cdot \frac{M}{N} - \frac{2N-1}{N} \cdot (\frac{1}{F} + \epsilon). \qquad (120)$$

Hence we have

$$R_{\text{ave}}^* \ge \max\left\{1 - \frac{M}{N}, \ \frac{2N-1}{N} - \frac{3N-2}{N} \cdot \frac{M}{N}\right\}$$
$$= R_{\text{u,ave}}(N, K, r). \tag{121}$$

#### Appendix I

## PROOF OF THEOREM 3 FOR AVERAGE RATE

To prove Theorem 3 for the average rate, we need to show that  $R_{ave}^* = R_u(N, K, r)$  for large N, for any caching system with no more than 5 users. Note that when N is large, with high probability all users will request distinct files. Hence, we only need to prove that the minimum average rate within the *type of the worst case demands* (i.e., the set of demands where all users request distinct files) equals  $R_u(N, K, r)$ . Since  $R_u(N, K, r)$  can already be achieved according to [24], it suffices to prove that this average rate is lower bounded by  $R_u(N, K, r)$ .

Similar to the peak rate case, we prove that this fact holds if  $\frac{KM}{N} \leq 1$  or  $\frac{KM}{N} \geq \lfloor \frac{K-3}{2} \rfloor$  for large N. When

 $\frac{KM}{N} \leq 1$  or  $\frac{KM}{N} \geq K - 1$ , this can be proved the same way as Lemma 4, while for the other case (i.e.  $\frac{KM}{N} \in [\max\{\lceil \frac{K-3}{2}\rceil, 1\}, K - 1)$ ), we need to prove a new version of Theorem 4, which lower bounds the average rate within the type of the worst case demands. To simplify the discussion, we adopt the notation of  $H^*(X_{\mathcal{D}}, W_{\mathcal{A}}, Z_{\mathcal{B}})$  which is defined in (89). We also adopt the corresponding notation for conditional entropy. Suppose rate R is  $\epsilon$  achievable for the worst case type, we start by proving converse bounds of R for large N.

Recall that  $r = \frac{KM}{N}$ , and let  $n = \lfloor r + 1 \rfloor$ . Because  $r \in [1, K - 1)$ , we have  $n \in \{2, ..., K - 1\}$ . Let  $\alpha = \lfloor \frac{N-K}{K-n} \rfloor$  and  $\beta = N - \alpha(K - n)$ . Suppose N is large enough, such that  $\alpha > 0$ . For any  $i \in \{1, ..., n\}$  and  $j \in \{1, ..., \alpha\}$  let  $d^{i,j}$  be a demand satisfying

$$d_{l}^{i,j} = \begin{cases} l - i + (j - 1)(K - n) + \beta \\ & \text{if } i + 1 \le l \le i + K - n, \\ l & \text{otherwise.} \end{cases}$$
(122)

Note that the above demands belong to the worst case type, so we have  $RF \ge H^*(X_{d^{i,j}})$  for any *i* and *j*. Following the same steps of proving Lemma 5, we have

$$RF \ge H^*(W_1|Z_1) + \frac{2}{n(n+1)\alpha} \left( \alpha n(K-n)F - nH^*(Z_1|W_{\{1,\dots,\beta\}}) - \sum_{i=0}^{K-n-1} H^*(Z_1|W_{\{1,\dots,\beta+i\alpha\}}) \right) - \frac{2K-n+1}{n+1}(1+\epsilon F).$$
(123)

Then following the steps of proving Theorem 4, we have

$$R \ge \frac{2K - n + 1}{n + 1} - \frac{K(K + 1)}{n(n + 1)} \cdot \frac{M}{N} - \frac{2K - n + 1}{n + 1}(\epsilon + \frac{1}{F})$$
(124)

if the following inequality holds:

$$K\beta + \alpha \frac{(K-n)(K-n-1)}{2} \le \frac{n(n+1)\alpha}{2}.$$
 (125)

Otherwise, we have

$$R \ge \frac{2K - n + 1}{n + 1} - \frac{2K(K - n)}{n(n + 1)} \cdot \frac{M}{N - \beta} - \frac{2K - n + 1}{n + 1} (\epsilon + \frac{1}{F}).$$
(126)

Similar to the proof of Lemma 4, we have proved that  $R \geq R_{\rm u}(N, K, r)$  from the above bounds if  $r \in \left[\max\{\lceil \frac{K-3}{2}\rceil, 1\}, K-1\right)$  for large N, large F, and small  $\epsilon$ . Consequently, we proved that  $R_{\rm ave}^* = R_{\rm u}(N, K, r)$  if  $r \leq 1$  or  $r \geq \lceil \frac{K-3}{2} \rceil$  for large N. For systems with no more than 5 users, this gives the exact characterization.

## $\begin{array}{c} \text{Appendix J} \\ \text{Convexity of } R_{\mathrm{u}}(N,K,r) \text{ and } R_{\mathrm{u,ave}}(N,K,r) \end{array}$

In this appendix, we prove the convexity of  $R_u(N, K, r)$  and  $R_{u,ave}(N, K, r)$  as functions of r, given parameters N and K. We start by proving the convexity of  $R_u(N, K, r)$ .

Recall that for any non-integer r, the value of  $R_u(N, K, r)$  is defined by linear interpolation. Hence, it suffices to show

that  $R_u(N, K, r)$  is convex on  $r \in \{0, 1, ..., K\}$ . Equivalently, we only need to prove

$$2R_{\rm u}(N,K,r) - R_{\rm u}(N,K,r-1) - R_{\rm u}(N,K,r+1) \le 0$$
(127)

for any  $r \in \{1, ..., K - 1\}$ .

The proof is as follows. We first observer that  $R_{\rm u}(N,K,r)$  can be written as

$$R_{\rm u}(N,K,r) = \frac{\binom{K}{r+1} - \binom{K-\min\{K,N\}}{r+1}}{\binom{K}{r}}$$
(128)  
$$\sum^{\min\{K,N\}} \frac{K-i}{K-i}$$

$$=\frac{\sum_{i=1}^{\min\{1,i,r\}}\binom{K-i}{r}}{\binom{K}{r}}$$
(129)

$$= \sum_{i=1}^{\min\{K,N\}} \frac{\binom{K-r}{i}}{\binom{K}{i}}.$$
 (130)

Consequently, the LHS of inequality (127) can be written as

$$2R_{u}(N, K, r) - R_{u}(N, K, r-1) - R_{u}(N, K, r+1)$$

$$=\sum_{i=1}^{\min\{K,N\}} \frac{2\binom{K-r}{i} - \binom{K-r-1}{i} - \binom{K-r+1}{i}}{\binom{K}{i}} \quad (131)$$

$$=\sum_{i=1}^{\min\{K,N\}} \frac{\binom{K-r-1}{i-1} - \binom{K-r}{i-1}}{\binom{K}{i}}$$
(132)

$$=\sum_{i=2}^{\min\{K,N\}} \frac{-\binom{K-r-1}{i-2}}{\binom{K}{i}}.$$
(133)

Since both  $\binom{K-r-1}{i-2}$  and  $\binom{K}{i}$  are non-negative, we have proved inequality (127). This guarantees the convexity of  $R_{\rm u}(N, K, r)$ .

Note that by substituting the variable  $\min\{K, N\}$  in function  $R_u(N, K, r)$  by  $N_e(d)$ , and taking expectation over a uniformly random demand d, we exactly obtain function  $R_{u,ave}(N, K, r)$ . Consequently, by applying the same substitution in the above proof, we obtain a proof for the convexity of  $R_{u,ave}(N, K, r)$ .

#### REFERENCES

- Q. Yu, M. A. Maddah-Ali, and A. S. Avestimehr, "Characterizing the rate-memory tradeoff in cache networks within a factor of 2," in 2017 *IEEE International Symposium on Information Theory (ISIT)*, pp. 386– 390, June 2017.
- [2] D. D. Sleator and R. E. Tarjan, "Amortized efficiency of list update and paging rules," *Communications of the ACM*, vol. 28, no. 2, pp. 202–208, 1985.
- [3] M. A. Maddah-Ali and U. Niesen, "Fundamental limits of caching," *IEEE Trans. Inf. Theory*, vol. 60, pp. 2856–2867, May 2014.
- [4] M. A. Maddah-Ali and U. Niesen, "Decentralized coded caching attains order-optimal memory-rate tradeoff," *IEEE/ACM Trans. Netw.*, vol. 23, pp. 1029–1040, Aug. 2015.
- [5] R. Pedarsani, M. A. Maddah-Ali, and U. Niesen, "Online coded caching," *IEEE/ACM Transactions on Networking*, vol. 24, no. 2, pp. 836–845, 2016.
- [6] U. Niesen and M. A. Maddah-Ali, "Coded caching with nonuniform demands," *IEEE Transactions on Information Theory*, 2016.
- [7] M. Ji, A. M. Tulino, J. Llorca, and G. Caire, "Order-optimal rate of caching and coded multicasting with random demands," *arXiv preprint arXiv:1502.03124*, 2015.
- [8] J. Zhang, X. Lin, and X. Wang, "Coded caching under arbitrary popularity distributions," in *Proc. ITA*, pp. 98–107, Feb. 2015.
- [9] M. Ji, G. Caire, and A. F. Molisch, "Fundamental limits of caching in wireless D2D networks," *IEEE Trans. Inf. Theory*, vol. 62, pp. 849–869, Feb. 2016.

- [10] S. H. Lim, C. Y. Wang, and M. Gastpar, "Information-theoretic caching: The multi-user case," *IEEE Transactions on Information Theory*, vol. 63, pp. 7018–7037, Nov 2017.
- [11] S. S. Bidokhti, M. Wigger, and R. Timo, "Noisy broadcast networks with receiver caching," arXiv preprint arXiv:1605.02317, 2016.
- [12] J. Zhang and P. Elia, "Fundamental limits of cache-aided wireless BC: Interplay of coded-caching and CSIT feedback," arXiv preprint arXiv:1511.03961, 2015.
- [13] J. Hachem, N. Karamchandani, and S. Diggavi, "Multi-level coded caching," in *Proc. IEEE ISIT*, pp. 56–60, 2014.
- [14] N. Karamchandani, U. Niesen, M. A. Maddah-Ali, and S. N. Diggavi, "Hierarchical coded caching," *IEEE Transactions on Information The*ory, vol. 62, no. 6, pp. 3212–3229, 2016.
- [15] M. A. Maddah-Ali and U. Niesen, "Cache-aided interference channels," in Proc. IEEE ISIT, pp. 809–813, June 2015.
- [16] N. Naderializadeh, M. A. Maddah-Ali, and A. S. Avestimehr, "Fundamental limits of cache-aided interference management," in *Proc. IEEE ISIT*, July 2016.
- [17] N. Naderializadeh, M. A. Maddah-Ali, and A. S. Avestimehr, "Cacheaided interference management in wireless cellular networks," in 2017 IEEE International Conference on Communications (ICC), pp. 1–7, May 2017.
- [18] Z. Chen, "Fundamental limits of caching: Improved bounds for small buffer users," arXiv preprint arXiv:1407.1935, 2014.
- [19] K. Wan, D. Tuninetti, and P. Piantanida, "On caching with more users than files," in *Information Theory (ISIT)*, 2016 IEEE International Symposium on, pp. 135–139, IEEE, 2016.
- [20] S. Sahraei and M. Gastpar, "K users caching two files: An improved achievable rate," in *Proc. CISS*, pp. 620–624, Mar. 2016.
- [21] C. Tian and J. Chen, "Caching and delivery via interference elimination," arXiv preprint arXiv:1604.08600, 2016.
- [22] M. M. Amiri and D. Gunduz, "Fundamental limits of caching: improved delivery rate-cache capacity trade-off," arXiv preprint arXiv:1604.03888, 2016.
- [23] M. M. Amiri, Q. Yang, and D. Gunduz, "Coded caching for a large number of users," arXiv preprint arXiv:1605.01993, 2016.
- [24] Q. Yu, M. A. Maddah-Ali, and A. S. Avestimehr, "The exact ratememory tradeoff for caching with uncoded prefetching," *IEEE Transactions on Information Theory*, vol. 64, pp. 1281–1296, Feb 2018.
- [25] J. Gómez-Vilardebó, "Fundamental limits of caching: Improved bounds with coded prefetching," arXiv preprint arXiv:1612.09071, 2016.
- [26] H. Ghasemi and A. Ramamoorthy, "Improved lower bounds for coded caching," *IEEE Transactions on Information Theory*, vol. 63, pp. 4388– 4413, July 2017.
- [27] A. Sengupta, R. Tandon, and T. C. Clancy, "Improved approximation of storage-rate tradeoff for caching via new outer bounds," in 2015 IEEE International Symposium on Information Theory (ISIT), pp. 1691–1695, IEEE, 2015.
- [28] C. Y. Wang, S. H. Lim, and M. Gastpar, "A new converse bound for coded caching," in 2016 Information Theory and Applications Workshop (ITA), pp. 1–6, Jan 2016.
- [29] C. Tian, "Symmetry, outer bounds, and code constructions: A computeraided investigation on the fundamental limits of caching," arXiv preprint arXiv:1611.00024, 2016.
- [30] N. S. Prem, V. M. Prabhakaran, R. Vaze, *et al.*, "Critical database size for effective caching," *arXiv preprint arXiv:1501.02549*, 2015.

#### BIOGRAPHIES

**Mohammad Ali Maddah-Ali** (S'03-M'08) received the B.Sc. degree from Isfahan University of Technology, and the M.A.Sc. degree from the University of Tehran, both in electrical engineering. From 2002 to 2007, he was with the Coding and Signal Transmission Laboratory (CST Lab), Department of Electrical and Computer Engineering, University of Waterloo, Canada, working toward the Ph.D. degree. From 2007 to 2008, he worked at the Wireless Technology Laboratories, Nortel Networks, Ottawa, ON, Canada. From 2008 to 2010, he was a post-doctoral fellow in the Department of Electrical Engineering and Computer Sciences at the University of California at Berkeley. Then, he joined Bell Labs, Holmdel, NJ, as a communication research scientist. Recently, he started working at Sharif University of Technology, as a faculty member.

Dr. Maddah-Ali is a recipient of NSERC Postdoctoral Fellowship in 2007, a best paper award from IEEE International Conference on Communications (ICC) in 2014, the IEEE Communications Society and IEEE Information Theory Society Joint Paper Award in 2015, and the IEEE Information Theory Society Joint Paper Award in 2016.

**A. Salman Avestimehr** (S'03-M'08-SM'17) is an Associate Professor at the Electrical Engineering Department of University of Southern California. He received his Ph.D. in 2008 and M.S. degree in 2005 in Electrical Engineering and Computer Science, both from the University of California, Berkeley. Prior to that, he obtained his B.S. in Electrical Engineering from Sharif University of Technology in 2003. His research interests include information theory, the theory of communications, and their applications to distributed computing and data analytics.

Dr. Avestimehr has received a number of awards, including the Communications Society and Information Theory Society Joint Paper Award, the Presidential Early Career Award for Scientists and Engineers (PECASE) for "pushing the frontiers of information theory through its extension to complex wireless information networks", the Young Investigator Program (YIP) award from the U. S. Air Force Office of Scientific Research, the National Science Foundation CAREER award, and the David J. Sakrison Memorial Prize. He is currently an Associate Editor for the IEEE Transactions on Information Theory.

**Qian Yu** (S'16) is pursuing his Ph.D. degree in Electrical Engineering at University of Southern California (USC), Viterbi School of Engineering. He received his M.Eng. degree in Electrical Engineering and B.S. degree in EECS and Physics, both from Massachusetts Institute of Technology (MIT). His interests span information theory, distributed computing, and many other problems math-related.

Qian is a recipient of the Google PhD Fellowship in 2018, and received the Jack Keil Wolf ISIT Student Paper Award in 2017. He received the Annenberg Graduate Fellowship in 2015, and Honorable Mention in the William Lowell Putnam Mathematical Competition in 2013.