

# Decentralized Caching and Coded Delivery with Distinct Cache Capacities

Mohammad Mohammadi Amiri, *Student Member, IEEE*, Qianqian Yang, *Student Member, IEEE*,  
and Deniz Gündüz, *Senior Member, IEEE*

**Abstract**—Decentralized proactive caching and coded delivery is studied in a content delivery network, where each user is equipped with a cache memory, not necessarily of equal capacity. Cache memories are filled in advance during the off-peak traffic period in a *decentralized* manner, i.e., without the knowledge of the number of active users, their identities, or their particular demands. User demands are revealed during the peak traffic period, and are served simultaneously through an error-free shared link. The goal is to find the minimum delivery rate during the peak traffic period that is sufficient to satisfy all possible demand combinations. A group-based decentralized caching and coded delivery scheme is proposed, and it is shown to improve upon the state-of-the-art in terms of the minimum required delivery rate when there are more users in the system than files. Numerical results indicate that the improvement is more significant as the cache capacities of the users become more skewed. A new lower bound on the delivery rate is also presented, which provides a tighter bound than the classical cut-set bound.

**Index Terms**—Coded caching, decentralized caching, distinct cache capacities, network coding, proactive caching.

## I. INTRODUCTION

The ever-increasing mobile data traffic is imposing a great challenge on the current network architectures. The growing demand has been typically addressed by increasing the achievable data rates; however, moving content to the network edge has recently emerged as a promising alternative solution as it reduces both the bandwidth requirements and the delay. In this paper, we consider an extreme form of edge caching, in which contents are stored directly at user terminals in a proactive manner. Proactive caching of popular contents, e.g., trending Youtube videos, episodes of popular TV series, during off-peak traffic periods also helps flattening the high temporal variability of traffic [1], [2].

In this proactive caching model [3], the *placement phase* takes place during off-peak traffic hours when the resources are abundant, without the knowledge of particular user demands. When the user demands are revealed, the *delivery phase* is performed, in which a common message is transmitted from the server to all the users over the shared communication channel. Each user decodes its requested file by combining the

bits received in the delivery phase with the contents stored in its local cache. Cache capacities are typically much lower than the size of the whole database, and a key challenge is to decide how to fill the cache memories without the knowledge of the user demands in order to minimize the *delivery rate*, which guarantees that all the user demands are satisfied, independent of the specific demand combination across the users. Maddah-Ali and Niesen showed in [3] that by storing and transmitting coded contents, and designing the placement and delivery phases jointly, it is possible to significantly reduce the delivery rate compared to uncoded caching.

A *centralized* caching scenario is studied in [3], in which the number and the identities of the users are known in advance by the server. This allows coordination of the cache contents across the users during the placement and delivery phases, such that by carefully placing pieces of contents in user caches a maximum number of multicasting opportunities are created to be exploited during the delivery phase. Many more recent works study centralized coded caching, and the required delivery rate has been further reduced [4]–[9].

In practice, however, the number or identities of active users that will participate in the *delivery phase* might not be known in advance during the *placement phase*. In such a scenario, called *decentralized caching*, coordination across users is not possible during the *placement phase*. For this scenario Maddah-Ali and Niesen proposed caching an equal number of random bits of each content at each user, and showed that one can still exploit multicasting opportunities in the delivery phase, albeit limited compared to the centralized counterpart [10]. *Decentralized caching* has been studied in various other settings, for example, with files with different popularities [11], [12], and distinct lengths [13], for online coded caching [14], coded caching of files with lossy reconstruction [15], as well as delivering contents from multiple servers over an interference channel [16], and in the presence of a fading delivery channel [17].

Most of the literature on coded caching assume identical cache capacities across users. However, in practice users access content through diverse devices, typically with very different storage capacities. Centralized caching with distinct cache capacities is studied in [15] and [18]. Recently, in [19], decentralized caching is studied for heterogeneous cache capacities; and by extending the scheme proposed in [10] to this scenario, authors have shown that significant gains can still be obtained compared to uncoded caching despite the asymmetry across users. In this paper, we propose a novel decentralized caching and delivery algorithm for users with

The authors are with Imperial College London, London SW7 2AZ, U.K. (e-mail: m.mohammadi-amiri15@imperial.ac.uk; q.yang14@imperial.ac.uk; d.gunduz@imperial.ac.uk).

This work received support from the European Union's H2020 Research and Innovation Programme through project TACTILENet: Towards Agile, effiCient, auTonomous and massIvely Large Network of things (agreement 690893), and from the European Research Council (ERC) through Starting Grant BEACON (agreement 677854).

distinct cache capacities. We show that the proposed scheme requires a smaller delivery rate than the one achieved in [19] when there are more users in the system than the number of files in the library, while the same performance is achieved otherwise. This scenario is relevant when a few popular video files or software updates are downloaded by many users within a short time period. Simulation results illustrate that the more distinct the cache capacities of the users are, which is more likely to happen in practice, the higher the improvement (with respect to [19]). We also derive an information-theoretic lower bound on the delivery rate building upon the lower bound derived in [20] for homogeneous cache capacities. This provides a lower bound on the delivery rate that is tighter than the classical cut-set bound.

The rest of this paper is organized as follows. The system model is introduced in Section II. In Section III, we present the proposed caching scheme as well as a lower bound on the delivery rate- cache capacity trade-off. The performance of the proposed coded caching scheme is compared with the state-of-the-art result analytically, and some numerical results are presented in Section IV. We conclude the paper in Section V. The detailed proofs are given in the Appendices.

*Notations:* The set of integers  $\{i, \dots, j\}$ , where  $i \leq j$ , is denoted by  $[i : j]$ . We denote the sequence  $Y_i, Y_{i+1}, \dots, Y_{j-1}, Y_j$  shortly by  $Y_{[i:j]}$ .  $\binom{j}{i}$  represents the binomial coefficient. For two sets  $\mathcal{Q}$  and  $\mathcal{P}$ ,  $\mathcal{Q} \setminus \mathcal{P}$  is the set of elements in  $\mathcal{Q}$  that do not belong to  $\mathcal{P}$ . Notation  $|\cdot|$  represents cardinality of a set, or the length of a file. Notation  $\oplus$  refers to bitwise XOR operation, while  $\oplus$  represents bitwise XOR operation where the arguments are first zero-padded to have the same length as the longest argument. Finally,  $\lfloor x \rfloor$  denotes the floor function; and  $(x)^+ \triangleq \max\{x, 0\}$ .

## II. SYSTEM MODEL

A server with a content library of  $N$  independent files  $W_{[1:N]}$  is considered. All the files in the library are assumed to be of length  $F$  bits, and each of them is chosen uniformly randomly over the set  $[1 : 2^F]$ . There are  $K$  active users,  $U_{[1:K]}$ , where  $U_k$  is equipped with a cache memory of capacity  $M_k F$  bits, with  $M_k < N$ ,  $\forall k^1$ . Data delivery is divided into two phases. User caches are filled during the *placement phase*. Let  $Z_k$  denote the contents of  $U_k$ 's cache at the end of the placement phase, which is a function of the database  $W_{[1:N]}$  given by  $Z_k = \phi_k(W_{[1:N]})$ , for  $k \in [1 : K]$ . Unlike in centralized caching [3], cache contents of each user are independent of the number and identities of other users in the system. User requests are revealed after the placement phase, where  $d_k \in [1 : N]$  denotes the demand of  $U_k$ , for  $k \in [1 : K]$ . These requests are served simultaneously through an error-free shared link in the *delivery phase*. The  $RF$ -bit message sent over the shared link by the server in response to the user demands  $d_{[1:K]}$  is denoted by  $X$ , where  $X \in [1 : 2^{RF}]$ , and it is generated by the encoding function  $\psi$ , i.e.,  $X = \psi(W_{[1:N]}, d_{[1:K]})$ .  $U_k$  reconstructs its requested file  $W_{d_k}$  after receiving the common message  $X$

<sup>1</sup>If  $M_k \geq N$ , for  $k \in [1 : K]$ ,  $U_k$  has enough memory to cache all the database; so,  $U_k$  does not need to participate in the delivery phase.

in the delivery phase along with its cache contents  $Z_k$ . The reconstruction at  $U_k$  for the demand combination  $d_{[1:K]}$  is given by  $\hat{W}_{d_k} = \rho_k(Z_k, X, d_{[1:K]})$ ,  $\forall k \in [1 : K]$ , where  $\rho_k$  is the decoding function at user  $U_k$ . For a given content delivery network, the tuple  $(\phi_{[1:K]}, \psi, \rho_{[1:K]})$  constitute a caching and delivery code with delivery rate  $R$ . We are interested in the *worst-case* delivery rate, that is the delivery rate that is sufficient to satisfy all demand combinations. Accordingly, the error probability is defined over all demand combinations as follows.

**Definition 1.** *The error probability of a  $(\phi_{[1:K]}, \psi, \rho_{[1:K]})$  caching and delivery code described above is given by*

$$P_e \triangleq \max_{d_{[1:K]} \in [1:N]^K} \Pr \left\{ \bigcup_{k=1}^K \{ \hat{W}_{d_k} \neq W_{d_k} \} \right\}. \quad (1)$$

**Definition 2.** *For a content delivery network with  $N$  files and  $K$  users, we say that a cache capacity-delivery rate tuple  $(M_{[1:K]}, R)$  is achievable if, for every  $\varepsilon > 0$ , there exists a caching and delivery code  $(\phi_{[1:K]}, \psi, \rho_{[1:K]})$  with error probability  $P_e < \varepsilon$ , for  $F$  large enough.*

There is a trade-off between the achievable delivery rate  $R$  and the cache capacities  $M_{[1:K]}$ , defined as

$$R^*(M_{[1:K]}) \triangleq \min \{ R : (M_{[1:K]}, R) \text{ is achievable} \}. \quad (2)$$

In this paper, we present upper and lower bounds on this trade-off.

## III. THE GROUP-BASED DECENTRALIZED CACHING (GBD) SCHEME

Here, we present the proposed *group-based decentralized (GBD) caching scheme*, first for uniform cache capacities, and then extend it to the scenario with distinct cache capacities.

### A. Uniform Cache Capacities

Here we assume that each user has the same cache capacity of  $MF$  bits, i.e.,  $M_1 = \dots = M_K = M$ .

**Placement phase:** In the placement phase, as in [10], each user caches a random subset of  $MF/N$  bits of each file independently. Since there are  $N$  files, each of length  $F$  bits, this placement phase satisfies the memory constraint.

For any set  $\mathcal{V} \subset [1 : K]$ ,  $W_{i,\mathcal{V}}$  denotes the bits of file  $W_i$  that have been *exclusively* cached by the users in set  $\mathcal{V}$ , that is,  $W_{i,\mathcal{V}} \subset Z_k$ ,  $\forall k \in \mathcal{V}$ , and  $W_{i,\mathcal{V}} \cap Z_k = \emptyset$ ,  $\forall k \in [1 : K] \setminus \mathcal{V}$ . For any chosen bit of a file, the probability of having been cached by any particular user is  $M/N$ . Since the contents are cached independently by each user, a bit of each file is cached exclusively by the users in set  $\mathcal{V} \subset [1 : K]$  (and no other user) with probability  $(M/N)^{|\mathcal{V}|} (1 - M/N)^{K-|\mathcal{V}|}$ .

**Delivery phase:** Without loss of generality, we order the users such that the first  $K_1$  users, referred to as group  $\mathcal{G}_1$ , demand  $W_1$ , the next  $K_2$  users, referred to as group  $\mathcal{G}_2$ , request  $W_2$ , and so on so forth. We define  $S_i \triangleq \sum_{l=1}^i K_l$ , which denotes the total number of users in the first  $i$  groups. Hence, the user demands are as follows:

$$d_k = i, \quad \text{for } i = 1, \dots, N, \text{ and } k = S_{i-1} + 1, \dots, S_i, \quad (3)$$

---

**Algorithm 1** Coded Delivery Phase for Uniform Cache Capacities Scenario
 

---

```

1: procedure CODED DELIVERY
2:   Part 1: Delivering bits that are not in the cache of any user
3:   for  $i = 1, \dots, N$  do
4:     send  $W_{d_{S_{i-1}+1}, \emptyset}$ 
5:   end for

6:   Part 2: Delivering bits that are in the cache of only one user
7:   send  $\bigcup_{i=1}^N \bigcup_{k=S_{i-1}+1}^{S_i-1} (W_{i,\{k\}} \oplus W_{i,\{k+1\}})$ 
8:   send  $\bigcup_{i=1}^{N-1} \bigcup_{j=i+1}^N \left( \bigcup_{k=S_{j-1}+1}^{S_j-1} (W_{i,\{k\}} \oplus W_{i,\{k+1\}}), \right.$ 
       $\left. \bigcup_{k=S_{i-1}+1}^{S_i-1} (W_{j,\{k\}} \oplus W_{j,\{k+1\}}), \right.$ 
       $\left. (W_{i,\{S_{j-1}+1\}} \oplus W_{j,\{S_{i-1}+1\}}) \right)$ 

9:   Part 3: Delivering bits that are in the cache of more than one user
10:  for  $\mathcal{V} \subset [1 : K] : 3 \leq |\mathcal{V}| \leq K$  do
11:    send  $\bigoplus_{v \in \mathcal{V}} W_{d_v, \mathcal{V} \setminus \{v\}}$ 
12:  end for
13: end procedure

14: procedure RANDOM DELIVERY [10]
15:  for  $i = 1, \dots, N$  do
16:    send sufficient random linear combinations (for reliable decoding) of the bits of file  $W_i$  to the users requesting it
17:  end for
18: end procedure

```

---

where we set  $S_0 = 0$ .

There are two alternative delivery procedures, called CODED DELIVERY and RANDOM DELIVERY, presented in Algorithm 1. The server follows the one that requires a smaller delivery rate. We present the CODED DELIVERY procedure of Algorithm 1 in detail, while we refer the reader to [10] for the RANDOM DELIVERY procedure, as we use the same procedure in [10] for the latter.

The main idea behind the CODED DELIVERY procedure is to deliver each user the missing bits of its request that have been cached by  $i$  user(s),  $\forall i \in [0 : K - 1]$ . In the first part, the bits of each request that are not in the cache of any user are directly delivered. Each transmitted content is destined for all the users in a distinct group, which have the same request.

In part 2, the bits of each request that have been cached by only one user are served. Note that, for any  $i \in [1 : N]$ , each user  $U_k$  in  $\mathcal{G}_i$ ,  $k \in [S_{i-1} + 1 : S_i]$ , demands  $W_i$  and has already cached  $W_{i,\{k\}}$ . Thus, having received the bits delivered in line 7 of Algorithm 1,  $U_k$  can recover  $W_{i,\{l\}}$ ,  $\forall l \in [S_{i-1} + 1 : S_i]$ , i.e., the bits of  $W_i$  cached by all the other users in the same group.

With the contents delivered in line 8 of Algorithm 1, each user can decode the subfiles of its requested file, which have been cached by users in other groups. Consider the users in two different groups  $\mathcal{G}_i$  and  $\mathcal{G}_j$ , for  $i = 1, \dots, N - 1$  and  $j = i + 1, \dots, N$ . All users in  $\mathcal{G}_i$  can recover subfile  $W_{j,\{S_i\}}$  after receiving  $\bigcup_{k=S_{i-1}+1}^{S_i-1} W_{j,\{k\}} \oplus W_{j,\{k+1\}}$ . Thus, they can obtain all subfiles  $W_{i,\{l\}}$ ,  $\forall l \in [S_{j-1} + 1 : S_j]$ , i.e., subfiles of  $W_i$  having been cached by users in  $\mathcal{G}_j$ , after receiving  $W_{i,\{S_{j-1}+1\}} \oplus W_{j,\{S_{i-1}+1\}}$  and  $\bigcup_{k=S_{j-1}+1}^{S_j-1} W_{i,\{k\}} \oplus W_{i,\{k+1\}}$ . Similarly, all users in  $\mathcal{G}_j$  can recover  $W_{i,\{S_j\}}$  after receiving  $\bigcup_{k=S_{j-1}+1}^{S_j-1} W_{i,\{k\}} \oplus W_{i,\{k+1\}}$ . Hence, by receiving  $W_{i,\{S_{j-1}+1\}} \oplus W_{j,\{S_{i-1}+1\}}$  and  $\bigcup_{k=S_{i-1}+1}^{S_i-1} W_{j,\{k\}} \oplus W_{j,\{k+1\}}$ , all users in  $\mathcal{G}_j$  can recover all the subfiles  $W_{j,\{l\}}$ ,  $\forall l \in [S_{i-1} + 1 : S_i]$ , i.e., subfiles of  $W_j$  that have been cached by users in  $\mathcal{G}_i$ .

In the last part, the same procedure as the one proposed in [10] is performed for the missing bits of each file that have been cached by more than one user. Consider any subset of users  $\mathcal{V} \subset [1 : K]$ , such that  $3 \leq |\mathcal{V}| \leq K$ . Each user  $v \in \mathcal{V}$  can recover subfile  $W_{d_v, \mathcal{V} \setminus \{v\}}$  after receiving the coded message delivered through line 11 of Algorithm 1. Hence, together with the local cache content and the contents delivered by the CODED DELIVERY procedure in Algorithm 1, each user can recover its desired file.

**Comparison with the state-of-the-art:** Here we compare the delivery rate of the proposed GBD scheme with that of the scheme proposed in [10, Algorithm 1] for uniform cache capacities scenario. The RANDOM DELIVERY procedure in Algorithm 1 is the same as the second delivery procedure of [10, Algorithm 1]. Thus, we focus only on the CODED DELIVERY procedure, and compare it with the first delivery procedure in [10, Algorithm 1]. The two procedures differ in the first and second parts. Consider a demand combination  $d_{[1:K]}$  with  $N'$  different requests, i.e.,

$$d_k = i, \quad \text{for } i = 1, \dots, N, \text{ and } k = S_{i-1} + 1, \dots, S_i, \quad (4)$$

such that  $S_i > 0$ , for  $i = 1, \dots, N'$ , and  $S_i = 0$ , for  $i = N' + 1, \dots, N$ . In the first part of CODED DELIVERY procedure in Algorithm 1, the bits of each  $N'$  different requested files, which have not been cached by any user, are delivered. A total of  $N'(1 - M/N)^K F$  bits are delivered in this part. On the other hand, in the first delivery procedure in [10, Algorithm 1], a total number of  $K(1 - M/N)^K F$  bits are delivered to serve the users with the bits which are not available in the cache of any user. From the fact that  $N' \leq \min\{N, K\}$  for any demand combination, the required number of bits delivered over the shared link in the first part of the CODED DELIVERY procedure in Algorithm 1 is smaller than or equal to that of the equivalent part in [10, Algorithm 1]. We further note that, if  $N < K$ , CODED DELIVERY procedure in Algorithm 1 delivers strictly less bits than [10, Algorithm 1] for this part of the delivery phase.

Next, we consider the second part of the CODED DE-

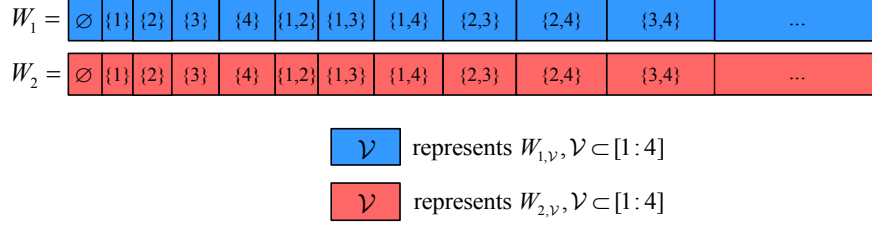


Fig. 1. Illustration of the subfiles, each corresponding to the bits of the file cached by a different subset of users.

LIVERY procedure. It is shown in Appendix A that a total of  $N'(K - (N' + 1)/2)$  coded contents, each of length  $(M/N)(1 - M/N)^{K-1}F$  bits, are delivered in the second part of the CODED DELIVERY procedure of the GBD scheme, leading to a delivery rate of

$$R_{\text{GBD}}^{\text{U}} \triangleq N' \left( K - \frac{N' + 1}{2} \right) \left( \frac{M}{N} \right) \left( 1 - \frac{M}{N} \right)^{K-1}. \quad (5)$$

On the other hand, the first delivery procedure in [10, Algorithm 1] sends a total of  $\binom{K}{2}$  coded contents in order to serve each user with the bits that have been cached by another user, leading to a delivery rate of

$$R_b^{\text{U}} \triangleq \frac{K(K-1)}{2} \left( \frac{M}{N} \right) \left( 1 - \frac{M}{N} \right)^{K-1}. \quad (6)$$

Since,  $N' \leq K$ , we have  $R_{\text{GBD}}^{\text{U}} \leq R_b^{\text{U}}$ , where the equality holds only if  $N' = K$  and  $N' = K - 1$ . Thus, for the case  $N < K - 1$ , which results in  $N' < K - 1$ , we have  $R_{\text{GBD}}^{\text{U}} < R_b^{\text{U}}$ , in which case the second part of the CODED DELIVERY procedure of the GBD scheme requires a smaller delivery rate than that of [10, Algorithm 1].

**Remark 1.** *The scheme proposed in [10] treats the users with the same demand as any other user, and it delivers the same number of bits for any demand combination; that is, for demand combination  $d_{[1:K]}$ , and any non-empty subset of users  $\mathcal{V} \subset [1:K]$ , it sends the coded content*

$$\bigoplus_{v \in \mathcal{V}} W_{d_v, \{\mathcal{V}\} \setminus \{v\}}, \quad (7)$$

*of length  $(M/N)^{|\mathcal{V}|-1}(1 - M/N)^{K-|\mathcal{V}|-1}$ , regardless of the redundancy among user demands. Instead, the proposed scheme treats the users with the same demand separately when delivering the bits of each file, and does not deliver redundant bits for the same demand.*

### B. Distinct Cache Capacities

In this section, we extend the proposed GBD scheme to the scenario with distinct cache capacities. We start with an illustrative example. It is then generalized to an arbitrary network setting. A new lower bound on the delivery rate is also obtained.

**Example 1.** Consider  $N = 2$  files,  $W_1$  and  $W_2$ , and  $K = 4$  users. Let the cache capacity of  $U_k$  be given by  $M_k = (1/2)^{4-k} M_{\text{max}}$ ,  $\forall k \in [1:4]$ .

In the placement phase,  $U_k$  caches a random subset of  $M_k F/2$  bits of each file independently. Since there are  $N = 2$

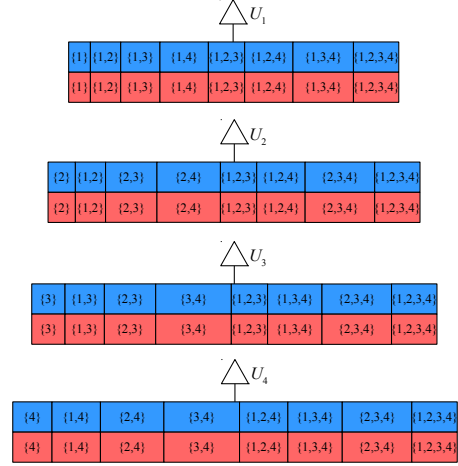


Fig. 2. Cache contents of users  $U_{[1:4]}$  after the placement phase.

files in the database, a total of  $M_k F$  bits are cached by  $U_k$ , filling up its cache memory. File  $W_i$  can be represented by

$$W_i = (W_{i,\mathcal{V}} : \forall \mathcal{V} \subset [1:4]), \quad \text{for } i = 1, 2. \quad (8)$$

An illustration of the subfiles, each cached by a different subset of users, is depicted in Fig. 1, and each user's cache content after the placement phase is shown in Fig. 2.

When  $N < K$ , it can be shown that the worst-case demand combination is the one when each of the  $N$  users with the smallest cache capacities requests a different file. For this particular example, we have  $M_1 \leq \dots \leq M_4$ , and accordingly, we have the worst-case demand combination when users  $U_1$  and  $U_2$ , i.e.,  $N = 2$  users with the smallest cache capacities, request distinct files. Hence, we can assume the worst-case demand combination of  $d_1 = d_3 = 1$  and  $d_2 = d_4 = 2$ .

As explained in Section III-A, the delivery phase consists of three distinct parts, where the bits delivered in part  $i$ ,  $i = 1, 2, 3$ , are denoted by  $X(i)$ , such that  $X = (X(1), X(2), X(3))$ . Below, we explain the purpose of each part in detail.

**Part 1:** In the first part of the delivery phase, the bits of each requested file which have not been cached by any user are directly delivered. For the example above, the following contents are delivered:  $X(1) = (W_{1,\emptyset}, W_{2,\emptyset})$ .

**Part 2:** The bits of the requested files, which have been cached exclusively by a single user (other than the requester) are sent in the second part of the delivery phase. The server first delivers each user the bits of

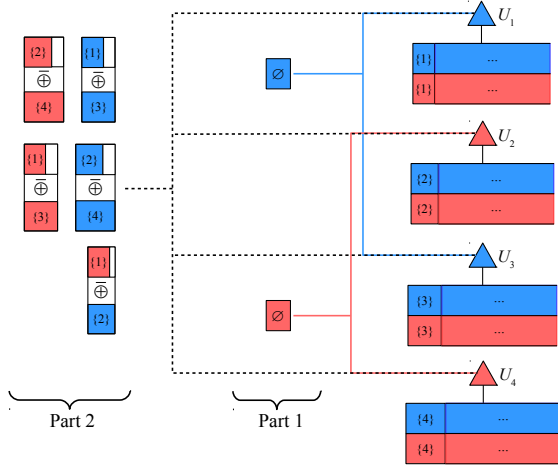


Fig. 3. Illustration of the coded contents delivered in parts 1 and 2 of the delivery phase for demand combination  $d_1 = d_3 = 1$  and  $d_2 = d_4 = 2$ , where the cached contents are shown in Fig. 2.

its requested file which are exclusively in the cache of one user with the same request. Then, each user receives the bits of its requested file which are only in the cache of a single user with a different request. In our example, with the following bits transmitted over the shared link,  $U_k$  can recover all the bits of its request  $W_{d_k}$ , which have been cached exclusively by  $U_l$ , for  $l \neq k$ ,  $k, l \in [1 : 4]$ :  $X(2) = (W_{1,\{3\}} \oplus W_{1,\{1\}}, W_{2,\{4\}} \oplus W_{2,\{2\}}, W_{1,\{4\}} \oplus W_{1,\{2\}}, W_{2,\{3\}} \oplus W_{2,\{1\}}, W_{1,\{2\}} \oplus W_{2,\{1\}})$ . The coded content delivered in parts 1 and 2 of the delivery phase have been illustrated in Fig. 3.

**Part 3:** In the last part, the server delivers the users the bits of their requested files which have been cached by more than one other user. Accordingly,  $U_k, \forall k \in [1 : 4]$ , can obtain all the bits of file  $W_{d_k}$ , which are in the cache of users in any set  $\mathcal{S} \subset [1 : 4] \setminus \{k\}$ , where  $|\mathcal{S}| \geq 2$ . For the example above, the following contents, illustrated in Fig. 4, are transmitted over the shared link:  $X(3) = (W_{1,\{2,3\}} \oplus W_{2,\{1,3\}} \oplus W_{1,\{1,2\}}, W_{1,\{2,4\}} \oplus W_{2,\{1,4\}} \oplus W_{2,\{1,2\}}, W_{1,\{3,4\}} \oplus W_{1,\{1,4\}} \oplus W_{2,\{1,3\}}, W_{2,\{3,4\}} \oplus W_{1,\{2,4\}} \oplus W_{2,\{2,3\}}, W_{1,\{2,3,4\}} \oplus W_{2,\{1,3,4\}} \oplus W_{1,\{1,2,4\}} \oplus W_{2,\{1,2,3\}})$ .

After receiving these three parts, each user can decode all the missing bits of its desired file. To find the delivery rate, we first note that, for  $F$  large enough, by the law of large numbers, the length of subfile  $W_{k,\mathcal{V}}$ , for any set  $\mathcal{V} \subset [1 : 4]$ , is approximately given by

$$|W_{k,\mathcal{V}}| \approx \prod_{i \in \mathcal{V}} \left( \frac{M_i}{2} \right) \prod_{j \in [1:4] \setminus \mathcal{V}} \left( 1 - \frac{M_j}{2} \right) F, \quad \forall k \in [1 : 4]. \quad (9)$$

Note that, due to the  $\oplus$  operation, the lengths of the delivered segments, e.g.,  $W_{1,\{2,3\}} \oplus W_{2,\{1,3\}} \oplus W_{1,\{1,2\}}$ , are given by the lengths of its longest arguments, i.e.,  $|W_{1,\{2,3\}} \oplus W_{2,\{1,3\}} \oplus W_{1,\{1,2\}}| = |W_{1,\{2,3\}}|$ . The delivery rate is given by the total rate of all the transmitted file segments listed above. When  $M_{\max} = 1$ , i.e.,  $M_{[1:K]} = (1/8, 1/4, 1/2, 1)$ , the delivery rate is 1.758, while the scheme in [19] would

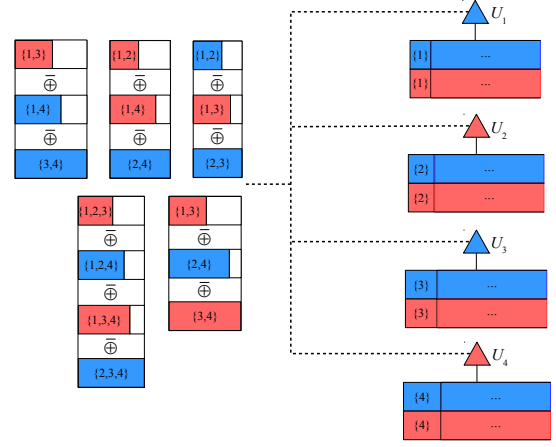


Fig. 4. Illustration of the coded contents delivered in part 3 of the delivery phase for demand combination  $d_1 = d_3 = 1$  and  $d_2 = d_4 = 2$ , where the cached contents are shown in Fig. 2.

require a delivery rate of 2.681. The GBD scheme provides a 34.43% reduction in the delivery rate compared to [19] in this example.  $\square$

Next, we present our caching and coded delivery scheme for the general case for arbitrary numbers of users and files, followed by the analysis of the corresponding delivery rate.

**Placement phase:** In the placement phase,  $U_k$  caches a random subset of  $M_k F/N$  bits of each file independently, for  $k = 1, \dots, K$ . Since there are  $N$  files in the database, a total of  $M_k F$  bits are cached by  $U_k$  satisfying the cache capacity constraint with equality. Since each user fills its cache independently, a bit of each file is cached exclusively by the users in set  $\mathcal{V} \subset [1 : K]$  (and no other user) with probability  $\prod_{i \in \mathcal{V}} (M_i/N) \prod_{j \in [1:K] \setminus \mathcal{V}} (1 - M_j/N)$ .

**Delivery phase:** We apply the same re-labeling of users into groups based on their requests as in Section III-A. We remind that the user demands are as follows:

$$d_k = i, \quad \text{for } i = 1, \dots, N, \text{ and } k = S_{i-1} + 1, \dots, S_i, \quad (10)$$

where users  $S_{i-1}, \dots, S_i$  form group  $\mathcal{G}_i$ . We further order the users within a group according to their cache capacities, and assume, without loss of generality, that  $M_{S_{i-1}+1} \leq M_{S_{i-1}+2} \leq \dots \leq M_{S_i}$ , for  $i = 1, \dots, N$ .

The delivery phase of the proposed GBD scheme for distinct cache capacities is presented in Algorithm 2. As in Section III-A, it has two distinct delivery procedures, CODED DELIVERY and RANDOM DELIVERY; and the server chooses the one with the smaller delivery rate.

The CODED DELIVERY procedure in Algorithm 2 follows the similar steps as the CODED DELIVERY procedure in Algorithm 1, except that  $\oplus$  is replaced with  $\oplus$ , due to the asymmetry across the users' cache capacities, and consequently, the size of the cached subfiles by different users. We remark that the correctness of the CODED DELIVERY in Algorithm 2 follows similarly to the correctness of the CODED DELIVERY procedure in Algorithm 1.

**Remark 2.** Note that  $W_{i,\{S_{j-1}+1\}}, \forall i, j \in [1 : N]$  such

---

**Algorithm 2** Coded Delivery Phase for Distinct Cache Capacities Scenario
 

---

```

1: procedure CODED DELIVERY
2:   Part 1: Delivering bits that are not in the cache of any
   user
3:   for  $i = 1, 2, \dots, N$  do
4:      $X(1) = (W_{d_{S_{i-1}+1}, \emptyset})$ 
5:   end for

6:   Part 2: Delivering bits that are in the cache of only
   one user
7:    $X(2, 1) = \left( \bigcup_{i=1}^N \bigcup_{k=S_{i-1}+1}^{S_i-1} (W_{i, \{k\}} \oplus W_{i, \{k+1\}}) \right)$ 
8:    $X(2, 2) = \bigcup_{i=1}^{N-1} \bigcup_{j=i+1}^N \left( \bigcup_{k=S_{j-1}+1}^{S_j-1} (W_{i, \{k\}} \oplus W_{i, \{k+1\}}), \right.$ 
    $\bigcup_{k=S_{i-1}+1}^{S_i-1} (W_{j, \{k\}} \oplus W_{j, \{k+1\}}),$ 
    $\left. (W_{i, \{S_{j-1}+1\}} \oplus W_{j, \{S_{i-1}+1\}}) \right)$ 

9:   Part 3: Delivering bits that are in the cache of more
   than one user
10:  for  $\mathcal{V} \subset [1 : K] : 3 \leq |\mathcal{V}| \leq K$  do
11:     $X(3) = \bigoplus_{v \in \mathcal{V}} W_{d_v, \mathcal{V} \setminus \{v\}}$ 
12:  end for
13: end procedure

14: procedure RANDOM DELIVERY
15:  for  $i = 1, 2, \dots, N$  do
16:    send enough random linear combinations of the
    bits of  $W_i$  to enable the users demanding it to decode it
17:  end for
18: end procedure

```

---

that  $i \neq j$ , is the smallest subfile of  $W_i$  cached exclusively by one user in  $\mathcal{G}_j$ . We also note that by sending any coded content  $W_{i, \{k_1\}} \oplus W_{j, \{k_2\}}$ ,  $k_1 \in [S_{j-1}+1 : S_j]$  and  $k_2 \in [S_{i-1}+1 : S_i]$ , instead of  $W_{i, \{S_{j-1}+1\}} \oplus W_{j, \{S_{i-1}+1\}}$  in  $X(2, 2)$ , for  $i = 1, \dots, N-1$  and  $j = i+1, \dots, N$ , the user demands can still be satisfied. However,  $W_{i, \{S_{j-1}+1\}} \oplus W_{j, \{S_{i-1}+1\}}$  has the smallest length among all coded contents  $W_{i, \{k_1\}} \oplus W_{j, \{k_2\}}$ ,  $\forall k_1 \in [S_{j-1}+1 : S_j]$  and  $\forall k_2 \in [S_{i-1}+1 : S_i]$ , which results in a smaller delivery rate.

In the RANDOM DELIVERY procedure, as in the second delivery procedure of [10, Algorithm 1], the server transmits enough random linear combinations of the bits of file  $W_i$  to the users in group  $\mathcal{G}_i$  such that they can all decode this file, for  $i = 1, \dots, N$ .

**Delivery rate analysis:** Consider first the case  $N \geq K$ . It can be argued in this case that the worst-case user demands happens if each file is requested by at most one user. Hence, by re-ordering the users, for the worst-case user demands, we have  $K_i = 1$ , for  $1 \leq i \leq K$ , and  $K_i = 0$ , otherwise. In this case, it can be shown that the CODED DELIVERY procedure requires a lower delivery

rate than the RANDOM DELIVERY procedure; hence, the server uses the former. In this case, it is possible to simplify the CODED DELIVERY procedure such that, only coded message  $X(2) = \bigcup_{i=1}^{N-1} \bigcup_{j=i+1}^N W_{i, \{S_{j-1}+1\}} \oplus W_{j, \{S_{i-1}+1\}}$  is transmitted in Part 2. The corresponding common message  $X = (X(1), X(2), X(3))$  transmitted over the CODED DELIVERY procedure reduces to the delivery phase of [19, Algorithm 2]. Thus, the GBD scheme achieves the same delivery rate as [19, Algorithm 2] when  $N \geq K$ .

Next, consider the case  $N < K$ . It is illustrated in Appendix B that the worst-case user demands happens when  $N$  users with the smallest cache capacities all request different files, i.e., they end up in different groups. The corresponding delivery rate is presented in the following theorem, the proof of which can also be found in Appendix B.

**Theorem 1.** *In a decentralized content delivery network with  $N$  files in the database, each of size  $F$  bits, and  $K$  users with cache capacities  $M_{[1:K]}$  satisfying  $M_1 \leq M_2 \leq \dots \leq M_K$ , the following delivery rate-cache capacity trade-off is achievable when  $N < K$ :*

$$R_{\text{GBD}}(M_{[1:K]}) = \min \left\{ \sum_{i=1}^K \prod_{j=1}^i \left( 1 - \frac{M_j}{N} \right) - \Delta R_1(M_{[1:K]}) - \Delta R_2(M_{[1:K]}) \right\}, \quad (11)$$

where

$$\Delta R_1(M_{[1:K]}) \triangleq (K - N) \prod_{l=1}^K \left( 1 - \frac{M_l}{N} \right), \quad (12a)$$

$$\Delta R_2(M_{[1:K]}) \triangleq \left[ \sum_{k=1}^{K-N} (k-1) \frac{M_{k+N}}{N - M_{k+N}} \right] \prod_{l=1}^K \left( 1 - \frac{M_l}{N} \right). \quad (12b)$$

**Comparison with the state-of-the-art:** Here the proposed GBD scheme for distinct cache capacities is compared with the scheme proposed in [19]. We note that, although the scheme presented in [19] is for  $N \geq K$ , it can also be applied to the case  $N < K$ , and the delivery rate given in [19, Theorem 2], denoted here by  $R_b(M_{[1:K]})$ , can still be achieved. Hence, in the following, when we refer to the scheme in [19, Algorithm 2], we consider its generalization to all  $N$  and  $K$  values. When  $N < K$ , according to [19, Theorem 2] and (11), we have

$$R_b(M_{[1:K]}) - R_{\text{GBD}}(M_{[1:K]}) \geq \Delta R_1(M_{[1:K]}) + \Delta R_2(M_{[1:K]}) > 0. \quad (13)$$

The second inequality in (13) holds as long as  $N < K$ . Therefore, when the number of files in the database is smaller than the number of active users in the delivery phase, the GBD scheme achieves a strictly smaller delivery rate than the one presented in [19].

**Remark 3.** *We note that the scheme of [19] exploits the caching scheme of [10] when the user cache capacities are distinct, and for any demand combination  $d_{[1:K]}$ , it delivers*

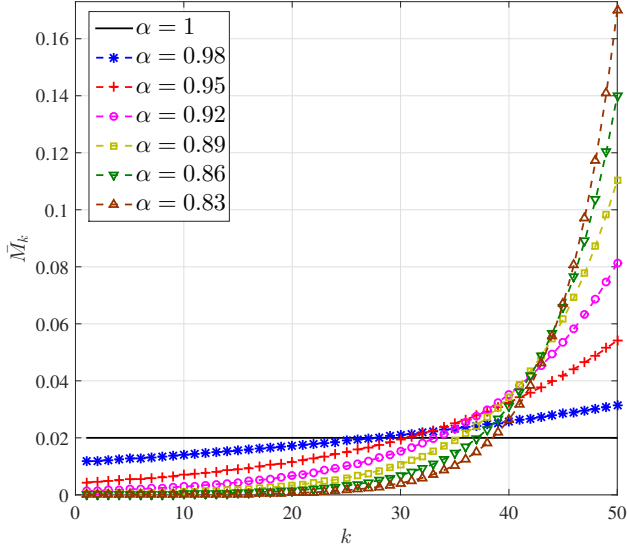


Fig. 5. Illustration of the normalized cache capacity distribution (normalized by  $\sum_{k=1}^K M_k$ ) for different  $\alpha$  values, in a cache network of  $K = 50$  users. The  $x$ -axis corresponds to the user index  $k$ .

$\bigoplus_{v \in \mathcal{V}} W_{d_v, \{\mathcal{V}\} \setminus \{v\}}$  to the users in any non-empty subset of users  $\mathcal{V} \subset [1 : K]$ , regardless of the users with the same demand. Thus, for the same reason explained in Remark 1 for uniform cache capacities, the proposed scheme in this paper outperforms the one in [19] for distinct cache capacities.

### C. Lower Bound on the Delivery Rate

In the next theorem, we generalize the information theoretic lower bound proposed in [20] to the content delivery network with distinct cache capacities. This lower bound is tighter than the classical cut-set bound.

**Theorem 2.** *In a content delivery network with  $N$  files in the database, serving  $K$  users with distinct cache capacities,  $M_{[1:K]}$  assorted in an ascending order, the optimal delivery rate satisfies*

$$R^*(M_{[1:K]}) \geq R_{LB}(M_{[1:K]}) = \max_{\substack{s \in [1:K], \\ l \in [1:\lceil N/s \rceil]}} \frac{1}{l} \times \left\{ N - \frac{s}{s+\gamma} \sum_{i=1}^{s+\gamma} M_i - \frac{\gamma(N-ls)^+}{s+\gamma} - (N-Kl)^+ \right\}, \quad (14)$$

where  $\gamma \triangleq \min \left\{ (\lceil N/l \rceil - s)^+, K - s \right\}, \forall s, l$ .

*Proof.* The proof of the theorem can be found in Appendix C.  $\square$

## IV. NUMERICAL RESULTS

In this section, the proposed GBD scheme for distinct cache capacities is compared with the scheme proposed in [19] numerically. To highlight the gains from the proposed scheme, we also evaluate the performance of uncoded caching, in which  $U_k, k \in [1 : K]$ , caches the first  $M_k/N$  bits of each file during the placement phase; and in the delivery phase the

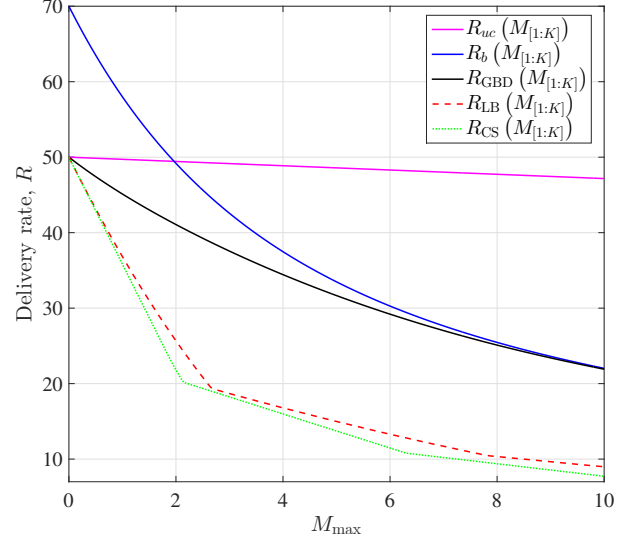


Fig. 6. Delivery rate versus  $M_{\max}$ , where the cache capacity of user  $k$  is  $M_k = \alpha^{K-k} M_{\max}, k = 1, \dots, K$ , with  $\alpha = 0.97, N = 50$ , and  $K = 70$ .

remaining  $1 - M_k/N$  bits of file  $W_{d_k}$  requested by  $U_k$  are delivered. By a simple analysis, it can be verified that the worst-case delivery rate is given by

$$R_{uc}(M_{[1:K]}) = \sum_{i=1}^{\min\{N, K\}} \left( 1 - \frac{M_i}{N} \right), \quad (15)$$

which is equal to the delivery rate of the RANDOM DELIVERY procedure in Algorithm 2.

For the numerical results, we consider an exponential cache capacity distribution among users, such that the cache capacity of  $U_k$  is given by  $M_k = \alpha^{K-k} M_{\max}$ , where  $0 \leq \alpha \leq 1$ , for  $k = 1, \dots, K$ , and  $M_{\max}$  denotes the maximum cache capacity in the system. Thus, we have  $M_{[1:K]} = (\alpha^{K-1} M_{\max}, \alpha^{K-2} M_{\max}, \dots, M_{\max})$ , which results in  $M_1 \leq M_2 \leq \dots \leq M_K$ , and the total cache capacity across the network is given by

$$\sum_{k=1}^K M_k = M_{\max} \alpha^k \frac{1 - \alpha}{1 - \alpha^{K+1}}. \quad (16)$$

The distribution of the cache capacities normalized by the total cache capacity available in the network, denoted by  $\bar{M}_k \triangleq M_k / \sum_{k=1}^K M_k, \forall k \in [1 : K]$ , is demonstrated in Fig. 5 for different values of  $\alpha$ , when  $K = 50$ . Observe that, the smaller the value of  $\alpha$ , the more skewed the cache capacity distribution across the users becomes. In the special case of  $\alpha = 1$ , we obtain the homogeneous cache capacity model studied in [10].

In Fig. 7, the delivery rate of the proposed scheme is compared with the scheme in [19] and uncoded caching, as well as the derived lower bound and the classical cut-set bound, when  $N = K = 3$  and  $\alpha = 0.8$ . The delivery rate is plotted with respect to the largest cache capacity in the system,  $M_{\max}$ . As expected, the delivery rate reduces as  $M_{\max}$  increases. This figure validates that the scheme proposed in Algorithm 2 achieves the same delivery rate as in [19] for

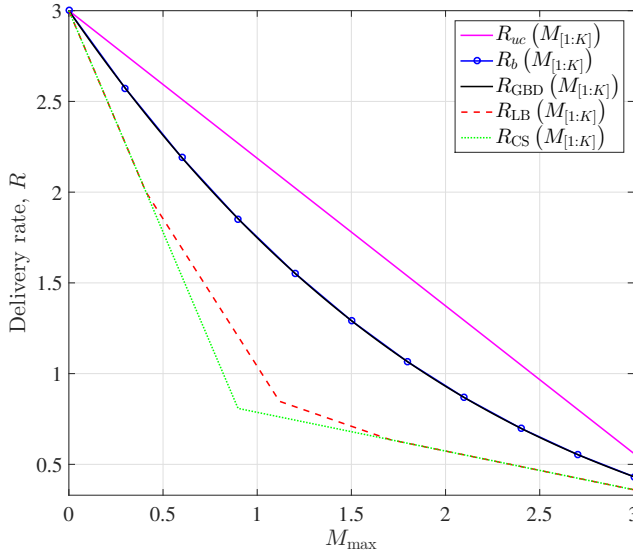


Fig. 7. Delivery rate versus  $M_{\max}$ , where the cache capacity of user  $k$  is  $M_k = \alpha^{K-k} M_{\max}$ ,  $k = 1, \dots, K$ , when  $N = K = 3$  and  $\alpha = 0.8$ .

$N \geq K$ . The GBD scheme achieves a significantly lower delivery rate compared to the uncoded scheme. It is to be noted that, the cut-set based lower bound derived in [19] is for the case  $N \geq K$ ; while, by ordering the users such that  $M_1 \leq M_2 \leq \dots \leq M_K$ , it can be re-written as follows for the general case:

$$R_{CS}(M_{[1:K]}) = \max_{s \in [1: \min\{N, K\}]} \left\{ s - \frac{\sum_{i=1}^s M_i}{\lfloor N/s \rfloor} \right\}. \quad (17)$$

The proposed lower bound is also plotted in Fig. 7. Similar to the case with identical cache sizes, the proposed lower bound is tighter than the cut-set lower bound for medium cache capacities. However, there remains a gap between this improved lower bound and the achievable delivery rate for the whole range of  $M_{\max}$  values.

In Fig. 6, the delivery rate  $R_{GBD}(M_{[1:K]})$  is compared with  $R_b(M_{[1:K]})$  and the uncoded scheme,  $R_{uc}(M_{[1:K]})$ , when  $N = 50$ ,  $K = 70$ , and  $\alpha = 0.97$ . We clearly observe that the proposed scheme outperforms both schemes at all values of  $M_{\max}$ . The improvement is particularly significant for lower values of  $M_{\max}$ , and it diminishes as  $M_{\max}$  increases. The proposed and the cut-set lower bounds are also included in the figure. Although the delivery rate of the proposed scheme meets the lower bounds when  $M_{\max} = 0$ , the gap in between quickly expands with  $M_{\max}$ .

In order to observe the effect of the skewness of the cache capacity distribution across users on the delivery rate, in Fig. 8, the delivery rate is plotted as a function of  $\alpha \in [0.9, 1]$ , for  $N = 30$ ,  $K = 45$ , and the largest cache capacity of  $M_{\max} = 2$ . Again, the GBD scheme achieves a lower delivery rate for the whole range of  $\alpha$  values under consideration. As opposed to uncoded caching, the gain over the scheme studied in [19] is more pronounced for smaller values of  $\alpha$ , i.e., as the distribution of cache capacities becomes more skewed. We

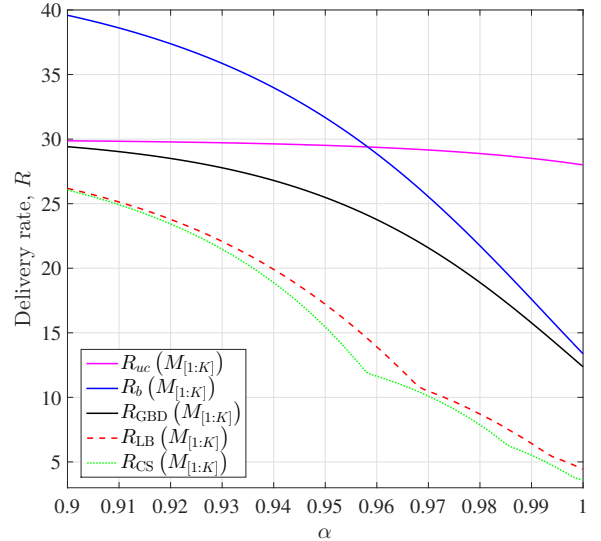


Fig. 8. Delivery rate versus  $\alpha \in [0.9, 1]$ , where  $M_k = \alpha^{K-k} M_{\max}$ , for  $k = 1, \dots, K$ , and  $N = 30$ ,  $K = 45$ , and  $M_{\max} = 2$ .

also observe that the gap to the lower bound is also smaller in this regime.

In Fig. 9, the delivery rate is plotted with respect to the number of users,  $K \in [1 : 100]$ , for  $N = 60$ ,  $M_{\max} = 5$ , and  $\alpha = 0.96$ . Observe that the improvement of the GBD scheme is more significant when the number of users requesting content in the delivery phase increases, whereas the gap between the GBD scheme and uncoded caching diminishes as  $K$  increases.

In Fig. 10, the delivery rate is plotted with respect to the number of files,  $N \in [10 : K - 1]$ , where the other parameters are fixed as  $K = 40$ ,  $M_{\max} = 4$ , and  $\alpha = 0.94$ . We observe that, the GBD scheme requires a smaller delivery rate compared to the state-of-the-art over the whole range of  $N$  values; while the improvement is more pronounced for smaller values of  $N$ . Observe also that, for relatively small values of  $N$ , the RANDOM DELIVERY procedure presented in Algorithm 2, which has the same performance as uncoded caching, outperforms the CODED DELIVERY procedure, i.e.,  $R_{RD}(M_{[1:K]}) < R_{CD}(M_{[1:K]})$ . The performance of uncoded caching gets worse with increasing  $N$  in this setting.

## V. CONCLUSIONS

We have studied proactive content caching at user terminals with distinct cache capacities, and proposed a novel caching scheme in a decentralized setting that improves upon the best known delivery rate in the literature. The improvement is achieved by creating more multicasting opportunities for the delivery of bits that have not been cached by any of the users, or cached by only a single user. In particular, the proposed scheme exploits the group-based coded caching scheme we have introduced previously for centralized content caching in a system with homogeneous cache capacities in [8]. Our numerical results show that the improvement upon the scheme proposed in [19] becomes more pronounced as the cache capacity distribution across users becomes more skewed,



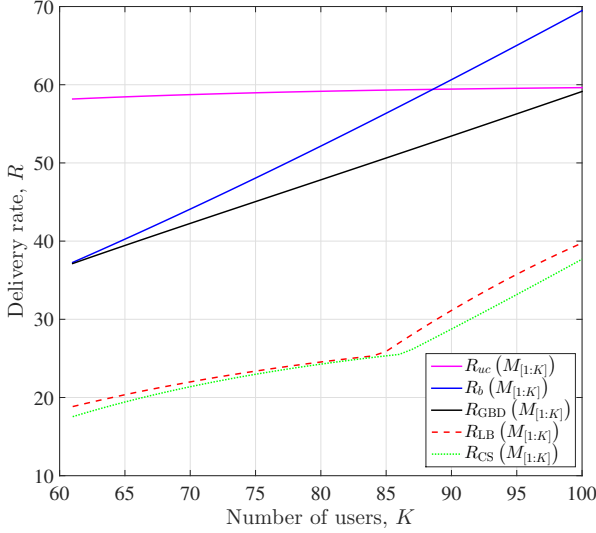


Fig. 9. Delivery rate versus the number of users  $K \in [1 : 100]$ , where  $M_k = \alpha^{K-k} M_{\max}$ , for  $k = 1, \dots, K$ , with  $N = 60$ ,  $M_{\max} = 5$ , and  $\alpha = 0.96$ .

showing that the proposed scheme is more robust against variations across user capabilities. We have also derived a lower bound on the delivery rate, which has been shown numerically to be tighter than the cut-set based lower bound studied in [19]. The gap between the lower bound and the best achievable delivery rate remains significant, calling for more research to tighten the gap in both directions.

#### APPENDIX A DELIVERY RATE IN PART 2 OF THE GBD SCHEME

If  $N' \leq \min\{N, K\}$  distinct files are requested by the users, without loss of generality, we order the users so that the users in  $\mathcal{G}_i$  request  $W_i$ , for  $i = 1, \dots, N'$ , i.e.,  $K_i = 0$ , for  $i = N' + 1, \dots, N$ . The coded contents delivered in line 7 of Algorithm 1 enable each user to obtain the subfiles of its requested file which are in the cache of one of the other users in the same group. Consider, for example, the first group, i.e.,  $i = 1$  in line 7 of Algorithm 1, which refers to the users that demand  $W_1$ . The XOR-ed contents  $W_{1,\{k\}} \oplus W_{1,\{k+1\}}$ , for  $k \in [1 : K_1 - 1]$ , are delivered by the server. Having subfile  $W_{1,\{k\}}$  cached, user  $U_k$ , for  $k \in [1 : K_1]$ , can decode all the remaining subfiles  $W_{1,\{j\}}$ , for  $j \in [1 : K_1] \setminus \{k\}$ . A total of  $(K_1 - 1)$  XOR-ed contents, each of size  $(M/N)(1 - M/N)^{K-1}F$  bits, are delivered for the users in  $\mathcal{G}_1$ . Similarly, for the second group ( $i = 2$  in line 7 of Algorithm 1), consisting of the users requesting  $W_2$ , the XOR-ed contents  $W_{2,\{k\}} \oplus W_{2,\{k+1\}}$ , for  $k \in [K_1 + 1 : K_1 + K_2 - 1]$ , are sent. With subfile  $W_{2,\{k\}}$  available locally at  $U_k$ , for  $k \in [K_1 + 1 : K_1 + K_2]$ ,  $U_k$  can obtain the missing subfiles  $W_{2,\{j\}}$ , for  $j \in [K_1 + 1 : K_1 + K_2] \setminus \{k\}$ . Hence, a total of  $(K_2 - 1)(M/N)(1 - M/N)^{K-1}F$  bits are served for the users in  $\mathcal{G}_2$ , and so on so forth. Accordingly,  $(K_i - 1)(M/N)(1 - M/N)^{K-1}F$  bits are delivered for group  $\mathcal{G}_i$ ,  $i = 1, \dots, N'$ , and the total number of bits sent by the server in the second

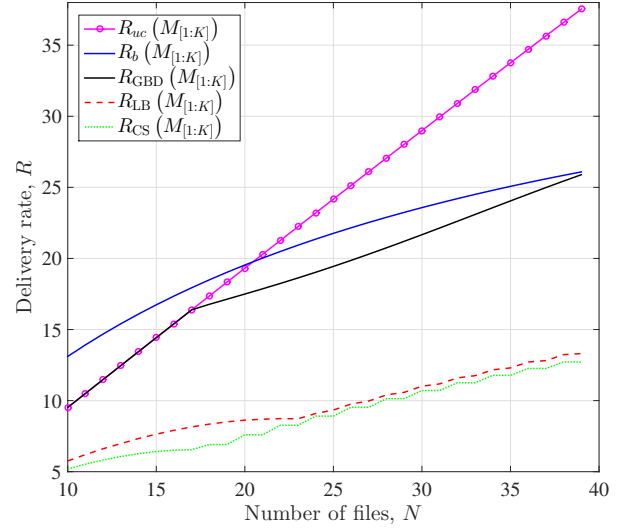


Fig. 10. Delivery rate versus  $N \in [10 : K - 1]$ , where  $M_k = \alpha^{K-k} M_{\max}$ , for  $k = 1, \dots, K$ , with  $K = 40$ ,  $M_{\max} = 4$ , and  $\alpha = 0.94$ .

part of the CODED DELIVERY procedure in Algorithm 1 is

$$\left(\frac{M}{N}\right) \left(1 - \frac{M}{N}\right)^{K-1} F \sum_{i=1}^{N'} (K_i - 1) = (K - N') \left(\frac{M}{N}\right) \left(1 - \frac{M}{N}\right)^{K-1} F. \quad (18)$$

After receiving the coded contents delivered in line 8 of Algorithm 1, each user in  $\mathcal{G}_i$ , for  $i \in [1 : N']$ , can recover the missing subfiles of its request  $W_i$ , which are in the cache of one of the users in groups  $j \in [1 : N'] \setminus \{i\}$ . Consider, for example,  $i = 1$  and  $j = 2$ . The XOR-ed contents  $W_{1,\{k\}} \oplus W_{1,\{k+1\}}$ , for  $k \in [K_1 + 1 : K_1 + K_2 - 1]$ , i.e., the subfiles of  $W_1$  cached by users in  $\mathcal{G}_2$ , are delivered. Next, the XOR-ed contents  $W_{2,\{k\}} \oplus W_{2,\{k+1\}}$ , for  $k \in [1 : K_1 - 1]$ , i.e., the subfiles of  $W_2$  cached by users in  $\mathcal{G}_1$ , are delivered. Finally, by delivering  $W_{1,\{K_1+K_2\}} \oplus W_{2,\{K_1\}}$ , and having already decoded  $W_{2,\{k\}}$  ( $W_{1,\{k\}}$ ),  $U_k$  in  $\mathcal{G}_1$  ( $\mathcal{G}_2$ ) can recover the missing subfiles of its request  $W_1$  ( $W_2$ ) which are in the cache of users in  $\mathcal{G}_2$  ( $\mathcal{G}_1$ ), for  $k \in [1 : K_1]$  (for  $k \in [K_1 + 1 : K_1 + K_2]$ ). The number of coded contents delivered in line 8 is  $(K_2 - 1) + (K_1 - 1) + 1$ , each of length  $(M/N)(1 - M/N)^{K-1}F$  bits, which adds up to a total number of  $(K_1 + K_2 - 1)(M/N)(1 - M/N)^{K-1}F$  bits. In a similar manner, the subfiles can be exchanged between users in groups  $\mathcal{G}_i$  and  $\mathcal{G}_j$ , for  $i \in [1 : N' - 1]$  and  $j \in [i + 1 : N']$ , by delivering a total of  $(K_i + K_j - 1)(M/N)(1 - M/N)^{K-1}F$  bits through sending the XOR-ed contents stated in line 8 of Algorithm 1. Hence, the total number of bits delivered in the second part of CODED DELIVERY procedure is given by

$$\left(\frac{M}{N}\right) \left(1 - \frac{M}{N}\right)^{K-1} F \sum_{i=1}^{N'-1} \sum_{j=i+1}^{N'} (K_i + K_j - 1) = (N' - 1) \left(K - \frac{N'}{2}\right) \left(\frac{M}{N}\right) \left(1 - \frac{M}{N}\right)^{K-1} F. \quad (19)$$

By summing up (18) and (19), the normalized number of

bits delivered in the second part of the CODED DELIVERY procedure in Algorithm 1 is given by

$$R_{\text{GBD}}^{\text{U}} = N' \left( K - \frac{N' + 1}{2} \right) \left( \frac{M}{N} \right) \left( 1 - \frac{M}{N} \right)^{K-1}. \quad (20)$$

#### APPENDIX B PROOF OF THEOREM 1

Consider first the CODED DELIVERY procedure in Algorithm 2. We note that, when  $N < K$ , the difference between the first procedure of the proposed delivery phase and the delivery phase presented in [19, Algorithm 1] lies in the first two parts, i.e., delivering the missing bits of the requested files, which either have not been cached by any user, or have been cached by only a single user. Hence, having the delivery rate of the scheme in [19, Algorithm 1], the delivery rate of the CODED DELIVERY procedure in Algorithm 2 can be determined by finding the difference in the delivery rates in these first two parts.

The delivery rate for Part 1 of the proposed CODED DELIVERY procedure, in which the bits of each request  $W_{d_k}$ , for  $k \in [1 : K]$ , that have not been cached by any user are directly sent to the users requesting the file, is given by

$$R_{\text{GBD}_1}(M_{[1:K]}) = N \prod_{k=1}^K \left( 1 - \frac{M_k}{N} \right). \quad (21)$$

We can see that the worst-case demand combination for this part of the CODED DELIVERY procedure is when each file is requested by at least one user, i.e.,  $K_i \geq 1, \forall i \in [1 : N]$ .

The corresponding delivery rate of [19, Algorithm 1] is given by:

$$R_{b_1}(M_{[1:K]}) = K \prod_{k=1}^K \left( 1 - \frac{M_k}{N} \right). \quad (22)$$

The difference between these two delivery rates is

$$\begin{aligned} \Delta R_1(M_{[1:K]}) &\triangleq R_{b_1}(M_{[1:K]}) - R_{\text{GBD}_1}(M_{[1:K]}) \\ &= (K - N) \prod_{k=1}^K \left( 1 - \frac{M_k}{N} \right). \end{aligned} \quad (23)$$

In Part 2 of the delivery phase of the GBD scheme, we deal with the bits of each requested file that have been cached by only a single user  $U_k$ , i.e.,  $W_{d_j, \{k\}}$ , for some  $k, j \in [1 : K]$ . For any request  $W_{d_j}$ , the normalized number of bits that have been cached exclusively by  $U_k$  will be denoted by  $Q_k$ . As  $F \rightarrow \infty$ , by the law of large numbers,  $Q_k$  can be approximated as [10]

$$\begin{aligned} Q_k &\approx \left( \frac{M_k}{N} \right) \prod_{l \in [1:K] \setminus \{k\}} \left( 1 - \frac{M_l}{N} \right) \\ &= \left( \frac{M_k}{N - M_k} \right) \prod_{l=1}^K \left( 1 - \frac{M_l}{N} \right). \end{aligned} \quad (24)$$

From (24) we can see that  $Q_i \geq Q_j, i \neq j, \forall i, j \in [1 : K]$ , if and only if  $M_i \geq M_j$ ; that is, the user with a larger cache size stores more bits of each file for  $F$  sufficiently large.

Next, we evaluate the delivery rate for Part 2 of the CODED DELIVERY procedure. We start with message  $X(2, 1)$ . For the users in  $\mathcal{G}_i$ , for  $i = 1, \dots, N$ , ordered in increasing cache capacities  $M_{S_{i-1}+1} \leq M_{S_{i-1}+2} \leq \dots \leq M_{S_i}$ , a total number of  $(K_i - 1)$  pieces, with the normalized sizes  $Q_{[S_{i-1}+2:S_i]}$ , are delivered. Thus, the delivery rate of the common message  $X(2, 1)$  is given by

$$R_{\text{GBD}_2}^1(M_{[1:K]}) \triangleq \sum_{i=1}^N \sum_{k=S_{i-1}+2}^{S_i} Q_k. \quad (25)$$

In line 8 of Algorithm 2,  $(K_j - 1)$  pieces, each of length  $Q_{[S_{j-1}+2:S_j]}$ , and  $(K_i - 1)$  pieces, each of length  $Q_{[S_{i-1}+2:S_i]}$  are delivered for users in  $\mathcal{G}_i$  and  $\mathcal{G}_j$ , respectively, for  $i = 1, \dots, N-1$  and  $j = i+1, \dots, N$ . Hence, the rate of the common message  $X(2, 2)$  is given by

$$R_{\text{GBD}_2}^2(M_{[1:K]}) \triangleq \sum_{i=1}^{N-1} \sum_{j=i+1}^N \left( \sum_{k=S_{j-1}+2}^{S_j} Q_k + \sum_{k=S_{i-1}+2}^{S_i} Q_k \right). \quad (26)$$

For each  $i \in [1 : N-1]$  and  $j \in [i+1 : N]$ , the normalized length of the bits delivered with the common message  $X(2, 3)$  is  $\max\{Q_{S_{j-1}+1}, Q_{S_{i-1}+1}\}$ . Thus, the rate of  $X(2, 3)$  is found to be:

$$R_{\text{GBD}_2}^3(M_{[1:K]}) \triangleq \sum_{i=1}^{N-1} \sum_{j=i+1}^N \max\{Q_{S_{j-1}+1}, Q_{S_{i-1}+1}\}. \quad (27)$$

To simplify the presentation, without loss of generality, let us assume that  $M_1 \leq M_{S_1+1} \leq \dots \leq M_{S_{N-1}+1}$ . Then (27) can be rewritten as

$$R_{\text{GBD}_2}^3(M_{[1:K]}) = \sum_{i=1}^{N-1} \sum_{j=i+1}^N Q_{S_{j-1}+1} = \sum_{k=1}^{N-1} k Q_{S_{k+1}}. \quad (28)$$

The total delivery rate for the second part of the proposed coded delivery phase is found by summing up the rates of the three parts, i.e.,

$$R_{\text{GBD}_2}(M_{[1:K]}) \triangleq \sum_{i=1}^3 R_{\text{GBD}_2}^i(M_{[1:K]}). \quad (29)$$

By substituting (25), (26), and (28) into (29), we get

$$R_{\text{GBD}_2}(M_{[1:K]}) = N \sum_{j=1}^N \sum_{k=S_{j-1}+2}^{S_j} Q_k + \sum_{k=1}^{N-1} k Q_{S_{k+1}}. \quad (30)$$

Note that, in (30), the coefficient of  $Q_{S_{k+1}}$  is  $k$ , for  $k \in [0 : N-1]$ , whereas the coefficient of all other  $Q_j$ s,  $\forall j \in [1 : K] \setminus \mathcal{P}$ , where  $\mathcal{P} \triangleq \{1, S_1+1, \dots, S_{N-1}+1\}$ , is  $N$ . Since  $N > K$ , the achievable rate for Part 2 of the CODED DELIVERY procedure in Algorithm 2 is maximized (the worst-case user demands happens) if  $Q_k \leq Q_j$ , for  $k \in \mathcal{P}$  and  $j \in [1 : K] \setminus \mathcal{P}$ ; or, equivalently, if  $M_k \leq M_j$ , for  $k \in \mathcal{P}$  and  $j \in [1 : K] \setminus \mathcal{P}$ . According to the definition of set  $\mathcal{P}$ , the above condition means that  $N$  users with the smallest cache sizes, i.e., users  $U_k, \forall k \in \mathcal{P}$ , will request different files, and belong to distinct groups in the worst-case scenario.

For simplification, without loss of generality, the users are ordered such that  $M_1 \leq M_2 \leq \dots \leq M_K$ . Then, the delivery rate of Part 2 of the CODED DELIVERY procedure is

$$R_{\text{GBD}_2}(M_{[1:K]}) = \sum_{k=1}^N (k-1) Q_k + N \sum_{k=N+1}^K Q_k. \quad (31)$$

By substituting  $Q_k$  in (24), we have

$$R_{\text{GBD}_2}(M_{[1:K]}) = \left[ \sum_{k=1}^N (k-1) \left( \frac{M_k}{N-M_k} \right) + N \sum_{k=N+1}^K \left( \frac{M_k}{N-M_k} \right) \right] \prod_{l=1}^K \left( 1 - \frac{M_l}{N} \right). \quad (32)$$

Now, we derive the delivery rate for the corresponding part in [19, Algorithm 1], i.e., when the server delivers the bits of the file requested by  $U_k$ , having been cached only by  $U_j$ ,  $\forall k, j \in [1:K]$ , such that  $j \neq k$ . For this case, from [19, Algorithm 1], when  $M_1 \leq M_2 \leq \dots \leq M_K$ , we have

$$R_{b_2}(M_{[1:K]}) = \left[ \sum_{k=1}^K (k-1) \left( \frac{M_k}{N-M_k} \right) \right] \prod_{l=1}^K \left( 1 - \frac{M_l}{N} \right). \quad (33)$$

Hence, the difference between the delivery rates for the second part of the proposed coded delivery phase and its counterpart in [19, Algorithm 1] is given by

$$\begin{aligned} \Delta R_2(M_{[1:K]}) &\triangleq R_{b_2}(M_{[1:K]}) - R_{\text{GBD}_2}(M_{[1:K]}) \\ &= \left[ \sum_{k=1}^{K-N} (k-1) \left( \frac{M_{k+N}}{N-M_{k+N}} \right) \right] \prod_{l=1}^K \left( 1 - \frac{M_l}{N} \right). \end{aligned} \quad (34)$$

Part 3 of the CODED DELIVERY procedure in Algorithm 2 is the same as its counterpart in [19, Algorithm 1]; so, they achieve the same delivery rate. Based on [19, Theorem 3], assuming that  $M_1 \leq M_2 \leq \dots \leq M_K$ , the delivery rate for the CODED DELIVERY procedure is

$$\begin{aligned} R_{\text{CD}}(M_{[1:K]}) &\triangleq \sum_{i=1}^K \left[ \prod_{j=1}^i \left( 1 - \frac{M_j}{N} \right) \right] \\ &\quad - \Delta R_1(M_{[1:K]}) - \Delta R_2(M_{[1:K]}), \end{aligned} \quad (35)$$

where  $\Delta R_1(M_{[1:K]})$  and  $\Delta R_2(M_{[1:K]})$  are as given in (23) and (34), respectively.

Now, consider the RANDOM DELIVERY procedure in Algorithm 2. Each delivered message in this procedure is directly targeted for the users in a group requesting the same file. It is assumed that the users in  $\mathcal{G}_i$  are ordered to have increasing cache capacities, such that  $M_{S_{i-1}+1} \leq M_{S_{i-1}+2} \leq \dots \leq M_{S_i}$ , for  $i = 1, \dots, N$ . Since each user in  $\mathcal{G}_i$  requires at most  $(1 - M_{S_{i-1}+1}/N)F$  bits to get its requested file, a total number of  $(1 - M_{S_{i-1}+1}/N)F$  bits, obtained from random linear combinations of  $W_i$ , are sufficient to enable the users in  $\mathcal{G}_i$  to decode their request  $W_i$ . Hence, the delivery rate for the RANDOM DELIVERY procedure in Algorithm 2 is

$$R_{\text{RD}}(M_{[1:K]}) \triangleq \sum_{i=1}^N \left( 1 - \frac{M_{S_{i-1}+1}}{N} \right). \quad (36)$$

Observe that the worst-case user demand combination corresponding to delivery rate  $R_{\text{RD}}(M_{[1:K]})$  happens (i.e., the delivery rate  $R_{\text{RD}}(M_{[1:K]})$  is maximized) when  $\{M_j, \forall j \in \mathcal{P}\}$  forms the set of  $N$  smallest cache capacities, i.e., the  $N$  users with the smallest cache capacities should request different files, which is consistent with the worst-case user demand combination corresponding to  $R_{\text{CD}}(M_{[1:K]})$ . If the users are labelled such that  $M_1 \leq M_2 \leq \dots \leq M_K$ , then we have

$$R_{\text{RD}}(M_{[1:K]}) = \sum_{i=1}^N \left( 1 - \frac{M_i}{N} \right). \quad (37)$$

We emphasize here that, before starting the *delivery phase*, it is assumed that each user sends its demand,  $d_k$ , together with its cache contents,  $Z_k$ , to the server. With this information, the server can perform the delivery procedure which requires a smaller delivery rate (by comparing (35) and (37)), and the following delivery rate is achievable:

$$R_{\text{GBD}}(M_{[1:K]}) \triangleq \min \{ R_{\text{CD}}(M_{[1:K]}), R_{\text{RD}}(M_{[1:K]}) \}, \quad (38)$$

which completes the proof of Theorem 1.

## APPENDIX C PROOF OF THEOREM 2

Our lower bound follows the techniques used in [20] to derive a lower bound for the setting with uniform cache capacities. For  $s \in [1:K]$ , it is assumed that the demands of the first  $s$  users are  $(d_{[1:s]}) = (1, \dots, s)$ , and the remaining  $(K-s)$  users have arbitrary demands  $d_k \in [1:N]$ ,  $\forall k \in [s+1:K]$ . The server delivers  $X_1 = \psi(W_{[1:N]}, 1, \dots, s, d_{[s+1:K]})$  to serve this demand combination. Now, consider the user demands  $(d_{[1:s]}) = (s+1, \dots, 2s)$ , and  $d_k \in [1:N]$ ,  $\forall k \in [s+1:K]$ , and the message  $X_2 = \psi(W_{[1:N]}, s+1, \dots, 2s, d_{[s+1:K]})$  delivered by the server to satisfy this demand combination. Consequently, considering the common messages  $X_{[1:\lceil N/s \rceil]}$  along with the cache contents  $Z_{[1:s]}$ , the whole database  $\{W_{[1:N]}\}$  can be recovered. We have

$$NF \leq H(Z_{[1:s]}, X_{[1:\lceil N/s \rceil]}) \quad (39a)$$

$$= H(Z_{[1:s]}) + H(X_{[1:\lceil N/s \rceil]} | Z_{[1:s]}) \quad (39b)$$

$$\leq \sum_{i=1}^s M_i F + H(X_{[1:\lceil N/s \rceil]} | Z_{[1:s]}) \quad (39c)$$

$$\begin{aligned} &= \sum_{i=1}^s M_i F + H(X_{[1:l]} | Z_{[1:s]}) \\ &\quad + H(X_{[l+1:\lceil N/s \rceil]} | Z_{[1:s]}, X_{[1:l]}) \end{aligned} \quad (39d)$$

$$\begin{aligned} &\leq \sum_{i=1}^s M_i F + lFR^*(M_{[1:K]}) \\ &\quad + H(X_{[l+1:\lceil N/s \rceil]} | Z_{[1:s]}, X_{[1:l]}) \end{aligned} \quad (39e)$$

$$\begin{aligned} &= \sum_{i=1}^s M_i F + lFR^*(M_{[1:K]}) \\ &\quad + H(X_{[l+1:\lceil N/s \rceil]} | Z_{[1:s]}, X_{[1:l]}, W_{[1:l]}) \\ &\quad + I(X_{[l+1:\lceil N/s \rceil]}; W_{[1:l]s} | Z_{[1:s]}, X_{[1:l]}) \end{aligned} \quad (39f)$$

$$\begin{aligned}
&\leq \sum_{i=1}^s M_i F + lFR^*(M_{[1:K]}) \\
&+ H(X_{[l+1:\lceil N/s \rceil]} | Z_{[1:s]}, X_{[1:l]}, W_{[1:ls]}) \\
&+ H(W_{[1:ls]} | Z_{[1:s]}, X_{[1:l]}) \quad (39g) \\
&\leq \sum_{i=1}^s M_i F + lFR^*(M_{[1:K]}) + \varepsilon l s F + 1 \\
&+ H(X_{[l+1:\lceil N/s \rceil]} | Z_{[1:s]}, X_{[1:l]}, W_{[1:ls]}) \quad (39h) \\
&\leq \sum_{i=1}^s M_i F + lFR^*(M_{[1:K]}) + \varepsilon l s F + 1 \\
&+ H(X_{[l+1:\lceil N/s \rceil]}, Z_{[s+1:s+\gamma]} | Z_{[1:s]}, X_{[1:l]}, W_{[1:ls]}) \quad (39i) \\
&= \sum_{i=1}^s M_i F + lFR^*(M_{[1:K]}) + \varepsilon l s F + 1 \\
&+ H(Z_{[s+1:s+\gamma]} | Z_{[1:s]}, X_{[1:l]}, W_{[1:ls]}) \\
&+ H(X_{[l+1:\lceil N/s \rceil]} | Z_{[1:s+\gamma]}, X_{[1:l]}, W_{[1:ls]}) \quad (39j) \\
&\leq \sum_{i=1}^s M_i F + lFR^*(M_{[1:K]}) + \varepsilon l s F + 1 \\
&+ H(Z_{[s+1:s+\gamma]} | Z_{[1:s]}, W_{[1:ls]}) \\
&+ H(X_{[l+1:\lceil N/s \rceil]} | Z_{[1:s+\gamma]}, X_{[1:l]}, W_{[1:ls]}) \quad (39k) \\
&= \sum_{i=1}^s M_i F + lFR^*(M_{[1:K]}) + \varepsilon l s F + 1 \\
&+ H(Z_{[1:s+\gamma]} | W_{[1:ls]}) - H(Z_{[1:s]} | W_{[1:ls]}) \\
&+ H(X_{[l+1:\lceil N/s \rceil]} | Z_{[1:s+\gamma]}, X_{[1:l]}, W_{[1:ls]}), \quad (39l)
\end{aligned}$$

where  $H(\cdot)$  denotes the entropy function, while  $I(\cdot; \cdot)$  represents the mutual information; (39d) follows from the chain rule of mutual information; (39e) follows from bounding the entropy of  $l$  common messages  $X_{[1:l]}$  given the cache contents  $Z_{[1:s]}$  by  $lFR^*(M_{[1:K]})$ ; (39f) is due to the definition of the mutual information; (39g) follows from the nonnegativity of entropy; (39h) is obtained from Fano's inequality; and (39i) also follows from the nonnegativity of entropy.

In (d),  $\gamma \leq K - s$  cache contents,  $Z_{[s+1:s+\gamma]}$ , are inserted inside the entropy. Note that from  $Z_{[s+1:s+\gamma]}$  together with messages  $X_{[1:l]}$  the remaining  $N - ls$  files in the database can be decoded. Since by each transmission  $X_i$  along with the caches  $Z_{[1:s+\gamma]}$ ,  $(s + \gamma)$  files can be decoded, we have  $s + \gamma \leq \lceil N/l \rceil$  for  $l$  number of transmissions. Hence, we have

$$\gamma = \min \left\{ \left( \left\lceil \frac{N}{l} \right\rceil - s \right)^+, K - s \right\}. \quad (40)$$

From the argument in [20, Appendix A], it can be verified that

$$H(X_{[l+1:\lceil N/s \rceil]} | Z_{[1:s+\gamma]}, X_{[1:l]}, W_{[1:ls]}) \leq (N - Kl)^+ F. \quad (41)$$

Based on (39) and (41), we have

$$\begin{aligned}
NF &\leq \sum_{i=1}^s M_i F - H(Z_{[1:s]} | W_{[1:ls]}) + lFR^*(M_{[1:K]}) \\
&+ H(Z_{[1:s+\gamma]} | W_{[1:ls]}) + (N - Kl)^+ F + \varepsilon l s F + 1. \quad (42)
\end{aligned}$$

Accordingly, for any set  $\mathcal{J} \subset [1 : s + \gamma]$  with  $|\mathcal{J}| = s$ , the

following inequality can be derived by choosing a set of caches  $\{Z_{\mathcal{J}}\} = \left\{ \bigcup_{k \in \mathcal{J}} Z_k \right\}$  that allows decoding the files in the database along with  $l$  common messages  $X_{[1:l]}$ :

$$\begin{aligned}
NF &\leq \sum_{i \in \mathcal{J}} M_i F - H(Z_{\mathcal{J}} | W_{[1:ls]}) + lFR^*(M_{[1:K]}) \\
&+ H(Z_{[1:s+\gamma]} | W_{[1:ls]}) + (N - Kl)^+ F + \varepsilon l s F + 1. \quad (43)
\end{aligned}$$

Hence, there are a total number of  $\binom{s+\gamma}{s}$  inequalities, each corresponding a different set  $\mathcal{J}$ . By taking average over all the inequalities, it can be evaluated that

$$\begin{aligned}
NF &\leq \frac{s}{s + \gamma} \sum_{i=1}^{s+\gamma} M_i F - \sum_{\substack{\mathcal{J} \subset [1:s+\gamma], \\ |\mathcal{J}|=s}} \frac{H(Z_{\mathcal{J}} | W_{[1:ls]})}{\binom{s+\gamma}{s}} \\
&+ lFR^*(M_{[1:K]}) + H(Z_{[1:s+\gamma]} | W_{[1:ls]}) \\
&+ (N - Kl)^+ F + \varepsilon l s F + 1. \quad (44)
\end{aligned}$$

By applying Han's inequality [21, Theorem 17.6.1], we have

$$\sum_{\substack{\mathcal{J} \subset [1:s+\gamma], \\ |\mathcal{J}|=s}} \frac{H(Z_{\mathcal{J}} | W_{[1:ls]})}{\binom{s+\gamma}{s}} \geq \frac{s}{s + \gamma} H(Z_{[1:s+\gamma]} | W_{[1:ls]}). \quad (45)$$

Accordingly, the following lower bound can be derived:

$$\begin{aligned}
NF &\leq \frac{s}{s + \gamma} \sum_{i=1}^{s+\gamma} M_i F + \frac{\gamma}{s + \gamma} H(Z_{[1:s+\gamma]} | W_{[1:ls]}) \\
&+ lFR^*(M_{[1:K]}) + (N - Kl)^+ F + \varepsilon l s F + 1. \quad (46)
\end{aligned}$$

It is shown in [20, Appendix A] that

$$H(Z_{[1:s+\gamma]} | W_{[1:ls]}) \leq (N - ls)^+ F. \quad (47)$$

From (46) and (47), we can obtain

$$\begin{aligned}
N &\leq \frac{s}{s + \gamma} \sum_{i=1}^{s+\gamma} M_i + \frac{\gamma(N - ls)^+}{s + \gamma} + lR^*(M_{[1:K]}) \\
&+ (N - Kl)^+ + \varepsilon l s + \frac{1}{F}. \quad (48)
\end{aligned}$$

For  $F$  large enough,  $\varepsilon > 0$  is arbitrary close to zero. As a result, we have

$$\begin{aligned}
R^*(M_{[1:K]}) &\geq \frac{1}{l} \times \\
&\left( N - \frac{s}{s + \gamma} \sum_{i=1}^{s+\gamma} M_i - \frac{\gamma(N - ls)^+}{s + \gamma} - (N - Kl)^+ \right). \quad (49)
\end{aligned}$$

By optimizing over all parameters  $s, l$ , and  $\gamma$ , and re-ordering the users such that  $M_1 \leq M_2 \leq \dots \leq M_K$  without loss of generality, we have

$$\begin{aligned}
R^*(M_{[1:K]}) &\geq R_{\text{LB}}(M_{[1:K]}) = \max_{\substack{s \in [1:K], \\ l \in [1:\lceil N/s \rceil]}} \frac{1}{l} \times \\
&\left\{ N - \frac{s}{s + \gamma} \sum_{i=1}^{s+\gamma} M_i - \frac{\gamma(N - ls)^+}{s + \gamma} - (N - Kl)^+ \right\}. \quad (50)
\end{aligned}$$

Note that, the first  $(s + \gamma)$  users have smaller cache capacities compared to all the other users. We can argue that the lower bound given in (50) is optimized over the set of cache capacities.

## REFERENCES

- [1] L. W. Dowdy and D. V. Foster, "Comparative models of the file assignment problem," *ACM Comput. Surv.*, vol. 14, pp. 287–313, Jun. 1982.
- [2] K. C. Almeroth and M. H. Ammar, "The use of multicast delivery to provide a scalable and interactive video-on-demand service," *IEEE J. Sel. Areas Commun.*, vol. 14, no. 6, pp. 1110–1122, Aug. 1996.
- [3] M. A. Maddah-Ali and U. Niesen, "Fundamental limits of caching," *IEEE Trans. Inform. Theory*, vol. 60, no. 5, pp. 2856–2867, May 2014.
- [4] Z. Chen, P. Fan, and K. B. Letaief, "Fundamental limits of caching: Improved bounds for users with small buffers," *IET Communications*, vol. 10, no. 17, pp. 2315–2318, Nov. 2016.
- [5] M. Mohammadi Amiri and D. Gündüz, "Improved delivery rate-cache capacity trade-off for centralized coded caching," in *Proc. IEEE ISITA*, Monterey, CA, Oct.-Nov. 2016, pp. 26–30.
- [6] —, "Fundamental limits of coded caching: Improved delivery rate-cache capacity trade-off," *IEEE Trans. Commun.*, vol. 65, no. 2, pp. 806–815, Feb. 2017.
- [7] C. Tian and J. Chen, "Caching and delivery via interference elimination," *arXiv:1604.08600v1 [cs.IT]*, Apr. 2016.
- [8] M. Mohammadi Amiri, Q. Yang, and D. Gündüz, "Coded caching for a large number of users," in *Proc. IEEE ITW*, Cambridge, UK, Sep. 2016, pp. 171–175.
- [9] J. Gomez-Vilardebo, "Fundamental limits of caching: Improved bounds with coded prefetching," *arXiv:1612.09071v4*, May 2017.
- [10] M. A. Maddah-Ali and U. Niesen, "Decentralized caching attains order-optimal memory-rate tradeoff," *IEEE/ACM Trans. Netw.*, vol. 23, no. 4, pp. 1029–1040, Apr. 2014.
- [11] U. Niesen and M. A. Maddah-Ali, "Coded caching with nonuniform demands," in *Proc. IEEE Conf. Comput. Commun. Workshops (INFOCOM WKSHPS)*, Toronto, ON, Apr. 2014, pp. 221–226.
- [12] M. Ji, A. M. Tulino, J. Llorca, and G. Caire, "Order-optimal rate of caching and coded multicasting with random demands," *arXiv:1502.03124v1 [cs.IT]*, Feb. 2015.
- [13] J. Zhang, X. Lin, C. C. Wang, and X. Wang, "Coded caching for files with distinct file sizes," in *Proc. IEEE Int'l Symp. on Inform. Theory*, Hong Kong, Jun. 2015, pp. 1686–1690.
- [14] R. Pedarsani, M. A. Maddah-Ali, and U. Niesen, "Online coded caching," in *Proc. IEEE Int. Conf. Commun. (ICC)*, Sydney, NSW, Jun. 2014, pp. 1878–1883.
- [15] Q. Yang and D. Gündüz, "Coded caching and content delivery with heterogeneous distortion requirements," *arXiv:1608.05660v1 [cs.IT]*, Aug. 2016.
- [16] J. P. Roig, D. Gündüz, and F. Tosato, "Interference networks with caches at both ends," in *Proc. IEEE Int'l Conf. on Communications (ICC)*, Paris, France, May 2017.
- [17] W. Huang, S. Wang, L. Ding, F. Yang, and W. Zhang, "The performance analysis of coded cache in wireless fading channel," *arXiv:1504.01452v1 [cs.IT]*, Apr. 2015.
- [18] A. Ibrahim, A. Zewail, and A. Yener, "Centralized coded caching with heterogeneous cache sizes," in *Proc. IEEE Wireless Communications and Networking Conference (WCNC)*, San Francisco, CA, Mar. 2017.
- [19] S. Wang, W. Li, X. Tian, and H. Liu, "Coded caching with heterogeneous cache sizes," *arXiv:1504.01123v3 [cs.IT]*, Aug. 2015.
- [20] A. Sengupta, R. Tandon, and T. C. Clancy, "Improved approximation of storage-rate tradeoff for caching via new outer bounds," in *Proc. IEEE Int'l Symp. on Inform. Theory*, Hong Kong, Jun. 2015, pp. 1691–1695.
- [21] T. M. Cover and J. A. Thomas, *Elements of Information Theory*. Hoboken, NJ, USA: Wiley-Interscience, John Wiley and Sons, Inc., 2006.



**Mohammad Mohammadi Amiri** (S'16) received the B.Sc. degree (Hons.) in communication systems from the Iran University of Science and Technology in 2011, and the M.Sc. degree (Hons.) in communication systems engineering from the University of Tehran in 2014. He is currently a Ph.D. candidate with Imperial College London under the supervision of Dr. Gündüz. His research interests include information and coding theory, wireless communications, MIMO systems, cooperative networks, cognitive radio, and signal processing.



**Qianqian Yang** (S'12) received the B.S. degree in automation from Chongqing University in 2011, and the M.S. degree in control engineering from Zhejiang University in 2014. She is currently a Ph.D. candidate with Imperial College London supervised by Dr. Gündüz. Her current research interests include information and coding theory, resource optimization in wireless sensor networks, neural network and machine learning.



**Deniz Gündüz** (S'03-M'08-SM'13) received the B.S. degree in electrical and electronics engineering from METU, Turkey in 2002, and the M.S. and Ph.D. degrees in electrical engineering from NYU Polytechnic School of Engineering in 2004 and 2007, respectively. After his PhD, he served as a postdoctoral research associate at Princeton University, and as a consulting assistant professor at Stanford University. He was a research associate at CTTC in Barcelona, Spain until September 2012, when he joined the Electrical and Electronic Engineering Department of Imperial College London, UK, where he is currently a Reader in information theory and communications.

His research interests lie in the areas of communications and information theory with special emphasis on multi-user communication networks, multimedia content delivery, energy efficient communications and privacy in cyber-physical systems. Dr. Gündüz is an Editor of the IEEE TRANSACTIONS ON COMMUNICATIONS, and the IEEE TRANSACTIONS ON GREEN COMMUNICATIONS AND NETWORKING. He is the recipient of a Starting Grant of the European Research Council (ERC) in 2016, IEEE Communications Society Best Young Researcher Award for the Europe, Middle East, and Africa Region in 2014, Best Paper Award at the 2016 IEEE Wireless Communications and Networking Conference (WCNC), and the Best Student Paper Award at the 2007 IEEE International Symposium on Information Theory (ISIT). He is the General Co-chair of the 2018 Workshop on Smart Antennas, and previously served as the General Co-chair of the 2016 IEEE Information Theory Workshop, and a Co-chair of the PHY and Fundamentals Track of the 2017 IEEE Wireless Communications and Networking Conference (WCNC).