Coded Caching with Heterogeneous File Demand Sets — The Insufficiency of Selfish Coded Caching

Chih-Hua Chang and Chih-Chun Wang
School of Electrical and Computer Engineering, Purdue University, U.S.A.

***Abstract*—This work falls under the broad setting of coded caching with user-dependent file popularity and average-rate capacity analysis. In general, the exact capacity characterization with user-dependent file popularity remains an open problem. For example, user 1 may be interested in files 1 and 2 with probabilities 0.6 and 0.4, respectively, while user 2 may be interested in only files 2, and 3 with probabilities 1/3 and 2/3, respectively, but not interested in file 1 at all. An optimal scheme needs to carefully balance the conflicting interests under the given probabilistic weights. Motivated by this fundamental but intrinsically difficult problem, this work studies the following simplified setting: Each user k is associated with a *file demand set* (FDS) Θ_k; each file in Θ_k is *equally desired* by user k with probability $\frac{1}{|\Theta_k|}$; and files outside Θ_k is not desired at all. Different users may have different $\Theta_{k_1} \neq \Theta_{k_2}$, which reflects the user-dependent file popularity. Various capacity results have been derived. One surprising byproduct is a proof showing that *selfish coded caching* is insufficient to achieve the capacity. That is, in an optimal coded caching scheme, a user sometimes has to cache the files of which he/she has zero interests.**

I. Introduction

Coded caching [1] could significantly reduce the worst-case peak-hour transmission time when compared to the traditional uncoded caching solutions. Existing works have characterized the coded caching capacity for some special N and K values [1]–[4] and derived order-optimal capacity expression for general N and K [1], [5]–[7].

While the focus on the worst-case performance is analytically appealing, it is oblivious to the underlying probability distribution of the random requests, and hence may not be able to address the phenomenon that the worst-case situation may only happen infrequently. Recently, there are new results focusing on the *average-rate capacity* [7]–[13]. Under the assumption that all users having the same file popularity profile[1] [7]–[11] proposed new order-optimal coded caching schemes and showed that the traditional (uncoded) highest-popularity-first policy can be strictly suboptimal. One justification of this *user-independent file popularity* is that we can put users of the same preference into a single group and perform coded caching within this group. Various other achievable average rate results have been proposed in [12], [13] under different levels of file/cache size heterogeneity but without the order optimality guarantee. They all assume the same user-independent file popularity setting.

Nonetheless, the above user-independent file popularity setting becomes less practical if each user has his/her own preference and the number of participating users is small, which is often the case in a traditional uncoded schemes that rely on highly individualized file prediction mechanisms. Motivated by the above observation, this work studies the average-rate capacity with user-dependent file popularity. The results would place coded caching on the same footing as the traditional uncoded solutions and allows for fair comparison between the two.

Specifically, we consider a coded caching system of N files and K users. Each user has his/her own file popularity of the N files. To reduce the complexity and the need of explicitly specifying K distinct probability distributions, one for each user, we assume that each user k is associated with a file demand set (FDS) Θ_k and is interested in those files in Θ_K with probability $\frac{1}{|\Theta_k|}$, see Section II for details. The FDS setting reflects user-dependent file popularity by allowing different users having distinct $\Theta_{k_1} \neq \Theta_{k_2}$. We then derive the average-rate capacity results under a variety of this FDS setting.

A byproduct of our capacity results is an answer to the following intuitive conjecture: *Since each user k is only interested in files in his/her own FDS Θ_k, there is zero incentive for user k to ever store any file that is outside Θ_k.* We prove that this conjecture holds for some very limited scenarios but is not true in general. This implies that for optimal coded caching, a user sometimes needs to encode beyond his/her own FDS, a surprising finding that defies the convention wisdom that one should only cache the files he/she is interested.

II. Problem Formulation

We consider a coded caching system with one server and K users. The server has access to N files $W_1, \ldots, W_N$, each having the same file size F bits. Each W_n is independently and uniformly randomly distributed over $\{0,1\}^F$. We use Z_k to denote the cache content of user k, which is of size M_k bits. Without loss of generality, we assume $M_k \in [0, NF]$. For any positive integer x, we define $[x] \triangleq \{1, \ldots, x\}$.

The operation of the system consists of the *placement phase* and the *delivery phase*. In the placement phase, each user k populates its cache content by

$$Z_k = \phi_k(W_1, \ldots, W_N),\ \forall k \in [K] \tag{1}$$

where ϕ_k is the caching function of user k. In the delivery phase, each user k sends a request $d_k \in [N]$ to the server,

This work was supported in parts by NSF under Grant CCF-1422997, Grant ECCS-1407604, Grant CCF-1618475, and Grant CCF-1816013.

[1]The popularity could be uniform across all N files [7] or vary significantly across all N files [8]–[10].

i.e., user k demands file W_{d_k}. We denote the probability mass function (pmf) of the random request d_k by $p^{[k]}_{d_k}$. The joint pmf of the demand pattern of K users $\vec{d} \triangleq (d_1, \ldots, d_K) \in [N]^K$ is then $p_{\vec{d}} = p^{[1]}_{d_1} \cdots p^{[K]}_{d_K}$.

After receiving the demand index vector $\vec{d}$, the server broadcasts an encoded signal

$$X_{\vec{d}} = \psi(\vec{d}, W_1, \ldots, W_N) \tag{2}$$

of $R_{\vec{d}}$ bits with encoding function ψ using an error-free link to all K users. Each user k then uses $X_{\vec{d}}$ as well as his/her cache content Z_k to decode the requested file

$$\hat{W}_{d_k} = \mu_k(\vec{d}, X_{\vec{d}}, Z_k), \tag{3}$$

where μ_k is the decoding function of user k. A coded caching scheme is completely specified by K caching functions $\{\phi_k\}$, one encoding function ψ, and K decoding functions $\{\mu_k\}$.

Definition 1. *The* file demand set *(FDS) of user k is defined as $\Theta_k \triangleq \{n \in [N] : p^{[k]}_n > 0\}$, which is the set of files that user k desires with a strictly positive probability.*

Definition 2. *A coded caching scheme is* zero-error feasible *if $\hat{W}_{d_k} = W_{d_k}$ for all $k \in [K]$, all $d_k \in \Theta_k$ in the FDS, and all $(W_1, \cdots, W_N) \in \{0,1\}^{NF}$.*

Throughout this manuscript, we consider exclusively zero-error feasible schemes.

Definition 3. *A coded caching scheme is* selfish *if we replace all K encoding functions ϕ_k in* (1) *by*

$$Z_k = \phi_k(\{W_n : n \in \Theta_k\}), \ \forall k \in [K]. \tag{4}$$

Namely, each user k only stores the files that he/she is interested, thus the name selfish*. In contrast, the original, more general design using* (1) *is referred to as an* unselfish *scheme.*

Definition 4. *The worst-case rate of a coded caching scheme is defined as*

$$R^* = \max_{\forall \vec{d}: d_k \in \Theta_k} R_{\vec{d}}. \tag{5}$$

Definition 5. *The average-rate of a coded caching scheme is defined as*

$$\bar{R} = \sum_{\forall \vec{d}, d_k \in \Theta_k} p_{\vec{d}} R_{\vec{d}}. \tag{6}$$

The uniform-average-rate of a scheme is defined as

$$\tilde{R} = \frac{1}{\prod_{k=1}^{K} |\Theta_k|} \sum_{\forall \vec{d}, d_k \in \Theta_k} R_{\vec{d}}. \tag{7}$$

$\tilde{R}$ can be viewed as a first-order approximation of the average-rate $\bar{R}$ that replaces the joint distribution $p_{\vec{d}}$ with a uniform distribution over the FDS $\prod_{k=1}^{K} \Theta_k$ (rather the simplest, uniform distribution over $[N]^K$ [7].). In [14], an exact characterization of $\bar{R}$ has been provided for the 2-user/2-file setting, which involves detailed discussion of up to 28 different cases that depends on the underlying values of (M_1, M_2) and $p_{\vec{d}}$. Instead of focusing on the exact $\bar{R}$, in this work we focus on the simplified, more tractable quantities $\tilde{R}$ and R^*, while relaxing the total number of files N being considered.

III. Main Results

Sections III-A and III-B present several exact capacity results. Then we present in Section III-C converse and achievability results that do not yet have a matching counterpart (thus not necessarily tight).

A. When Selfish and Unselfish Designs Are Equally Powerful

In this subsection, we outline several special cases for which selfish and unselfish designs are equally powerful.

Proposition 1. *If $\Theta_{k_1} \cap \Theta_{k_2} = \emptyset$ for all distinct $k_1, k_2 \in [K]$, then selfish and unselfish designs are equally powerful and achieve the same R^* and $\tilde{R}$.*

The proof of Proposition 1 is delegated in Appendix A. This proposition shows if no two users are interested in a common file, each user can act as if he/she is the sole user in the system.

Proposition 2 ($\Theta_1 = \Theta_2$)**.** *Consider $K = 2$ users, $N \geq 2$ files, and $\Theta_1 = \Theta_2 = [N]$. By definition, there is no difference between selfish and unselfish designs. Then the $\tilde{R}$ is tightly characterized*[2] *by*

$$\tilde{R} \geq F - (M_1/N) \tag{Q1}$$

$$\tilde{R} \geq F - (M_2/N) \tag{Q2}$$

$$\tilde{R} \geq \frac{2N-1}{N}F - \frac{2N-2}{N^2}M_1 - \frac{1}{N}M_2 \tag{Q3}$$

$$\tilde{R} \geq \frac{2N-1}{N}F - \frac{1}{N}M_1 - \frac{2N-2}{N^2}M_2 \tag{Q4}$$

The proof of Proposition 2 is delegated in Appendix B.

This proposition is the average-rate counterpart of the worst-case setting in [3] for the $K = 2$ and arbitrary $N \geq 2$ case. The relationship of $\tilde{R}$ versus (M_1, M_2) is illustrated in Fig. 1. The x-axis (resp. y-axis) is for the M_1 (resp. M_2) value. The inequalities (Q1) to (Q4) are marked in the corresponding regions. There are seven vertices t_1 to t_7 and each vertex is labeled by a tuple $(M_1, M_2, \tilde{R})$, where (M_1, M_2) describe the location and the third coordinate describe the corresponding exact uniform-average-rate capacity $\tilde{R}$.

We then consider the simplest scenario when $\{\Theta_k\}$ are distinct and overlap with each other.

Proposition 3 ($\Theta_1 = \{1\}, \Theta_2 = [N]$)**.** *Consider $K = 2$ users and $\Theta_1 = \{1\}$ and $\Theta_2 = [N]$. The selfish and unselfish designs have identical $\tilde{R}$, which is tightly characterized by:*

$$\tilde{R} \geq F - M_1 \tag{Q5}$$

$$\tilde{R} \geq F - (M_2/N) \tag{Q6}$$

$$\tilde{R} \geq \frac{2N-1}{N}F - \frac{N-1}{N}M_1 - \frac{1}{N}M_2 \tag{Q7}$$

The proof of Proposition 3 is delegated in Appendix C.

The relationship of $\tilde{R}$ versus (M_1, M_2) is illustrated in Fig. 2. Comparing Propositions 2 and 3, it is clear that when the FDS Θ_1 reduces from $[N]$ to $\{1\}$, the capacity $\tilde{R}$ reduces since an optimal scheme can now take advantage of the fact that user 1 is only interested in file 1.

[2]We use the statement *tightly characterized* when we can derive a matching pair of the converse and achievability results, i.e., it characterizs capacity.

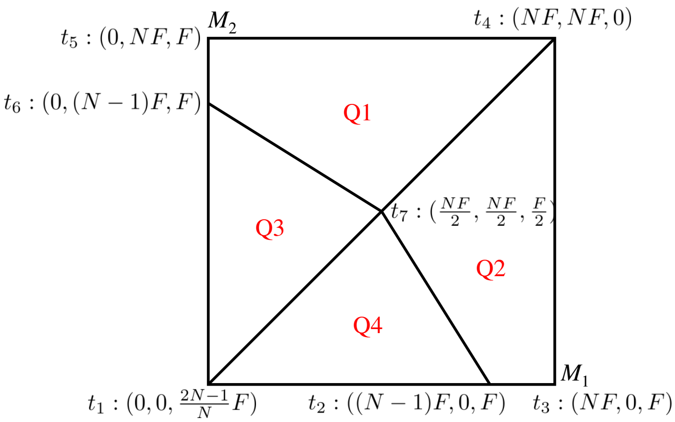

Fig. 1. The capacity $\tilde{R}$ of both the selfish and unselfish designs w. $\Theta_1 = \Theta_2 = [N]$.

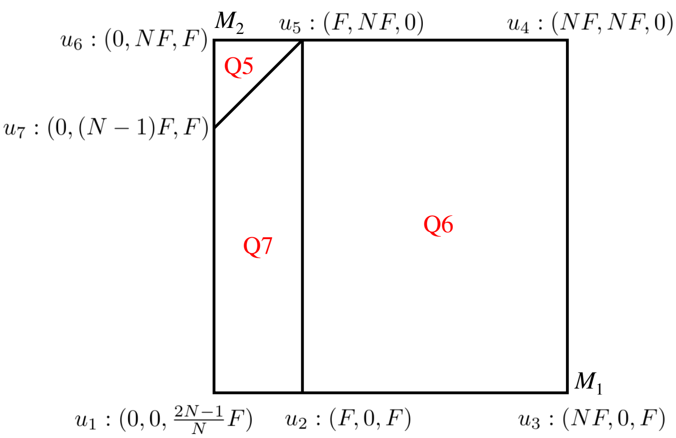

Fig. 2. The capacity $\tilde{R}$ of both the selfish and unselfish designs w. $\Theta_1 = \{1\}$ and $\Theta_2 = [N]$.

B. Insufficiency of Selfish Designs

Proposition 4 (Unselfish w. $\Theta_1 = \{1,2\}, \Theta_2 = \{1,2,3\}$)**.** *Consider $K = 2$ users and $\Theta_1 = \{1,2\}$ and $\Theta_2 = \{1,2,3\}$. $\tilde{R}$ of the unselfish schemes is tightly characterized by*

$$\tilde{R} \geq F - M_1/2 \tag{P1}$$

$$\tilde{R} \geq F - M_2/3 \tag{P2}$$

$$\tilde{R} \geq \frac{5F}{4} - \frac{M_1}{4} - \frac{M_2}{4} \tag{P3}$$

$$\tilde{R} \geq \frac{3F}{2} - \frac{M_1}{3} - \frac{M_2}{3} \tag{P4}$$

$$\tilde{R} \geq \frac{5F}{3} - \frac{M_1}{2} - \frac{M_2}{3} \tag{P5}$$

$$\tilde{R} \geq \frac{5F}{3} - \frac{M_1}{3} - \frac{M_2}{2} \tag{P6}$$

The proof of Proposition 4 is provided in Appendix D. The relationship of the unselfish capacity $\tilde{R}$ versus (M_1, M_2) is illustrated in Fig. 3.

Proposition 5 (Selfish w. $\Theta_1 = \{1,2\}, \Theta_2 = \{1,2,3\}$)**.** *Continue from Proposition 4. $\tilde{R}$ of the* selfish schemes *is tightly characterized by* (P1) *to* (P6) *plus an additional inequality:*

$$\tilde{R} \geq \frac{4F}{3} - \frac{M_1}{6} - \frac{M_2}{3}. \tag{P7}$$

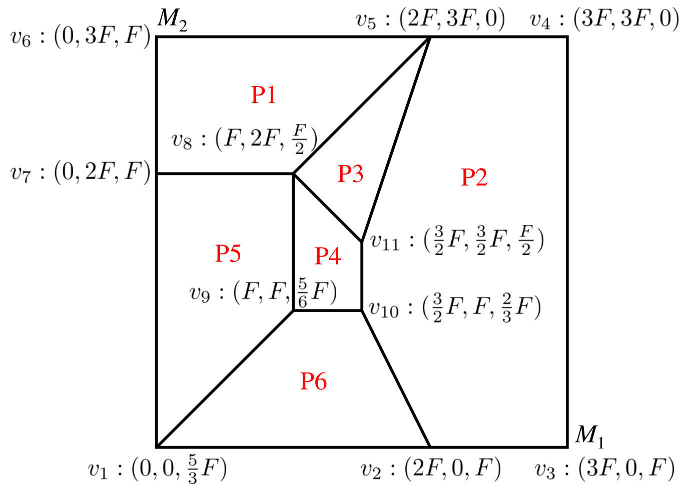

Fig. 3. The unselfish capacity $\tilde{R}$ w. $\Theta_1 = \{1,2\}$ and $\Theta_2 = \{1,2,3\}$.

The proof of Proposition 5 is provided in Appendix E.

The relationship of the selfish capacity $\tilde{R}$ versus (M_1, M_2) is illustrated in Fig. 4. Note that one can prove that if $\tilde{R}$ satisfies inequality (P7), then it automatically satisfies (P3) and (P4). That is why in Fig. 4 there are only 5 subregions and the regions of (P3) and (P4) no longer appear.

When viewed separately, Propositions 4 and 5 describe the fundamental limits of unselfish and selfish coded caching when two users, with arbitrary cache sizes (M_1, M_2), share concentrated[3], similar, but not identical interests, which alone are of important analytical value. Jointly, they provide the first proof that selfish coded caching is strictly suboptimal, e.g., the two points v_{10} and v_{11} in Fig. 3 can only be achieved by an unselfish design.

It is worth pointing out that the insufficiency of selfish coded caching is not due to the use of the average rate $\tilde{R}$ as the performance metric. Even when using the worst-case rate R^* in (5), selfish designs are still insufficient.

Corollary 1. *Continue from Proposition 4. When $(M_1, M_2) = (1.5F, 1.5F)$, i.e., v_{11} in Fig. 3, the worst-case capacity R^* of the unselfish and selfish schemes are $0.5F$ and $\frac{7}{12}F$, respectively.*

By Propositions 4 and 5, we quickly see that the average-rate capacity $\tilde{R}$ of the unselfish and selfish schemes are $0.5F$ and $\frac{7}{12}F$, respectively. We then show that the worst-case rate $\frac{7}{12}F$ is achievable. We denote the three files by (A, B, C) such that user 1 demands files (A, B) and user 2 demands files (A, B, C). We divide $A = (A_1, A_2, A_3, A_4)$, $B = (B_1, B_2, B_3, B_4)$, $C = (C_1, C_2)$ into disjoint subfiles, where A_1, B_1 has size $\frac{2}{12}F$; A_2 and B_2 has size $\frac{4}{12}F$; A_3, A_4, B_3, and B_4 has size $\frac{3}{12}F$; C_1 has size $\frac{2}{3}F$; and

[3] We say a user is of *concentrated interest* if the corresponding FDS Θ_k is small, e.g., $|\Theta_1| = 2$ and $|\Theta_2| = 3$ in Propositions 4 and 5. This is usually a result of highly effective next-file prediction.

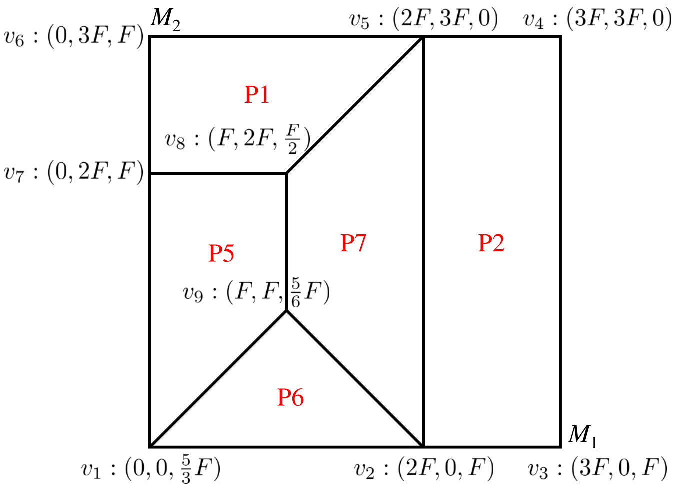

Fig. 4. The selfish capacity $\tilde{R}$ w. $\Theta_1 = \{1, 2\}$ and $\Theta_2 = \{1, 2, 3\}$.

C_2 has size $\frac{1}{3}F$. In the placement phase, user 1 caches $Z_1 = (A_1, A_2, A_3, B_1, B_2, B_3)$ of size $\frac{3}{2}F$ and user 2 caches $Z_2 = (A_1, A_4, B_1, B_4, C_1)$ of size $\frac{3}{2}F$. In the delivery phase, the transmitted signal and the rate of 6 demands are

(d_1, d_2)	$X_{(d_1,d_2)}$	$R_{(d_1,d_2)}$
$(1, 1)$	$(A_2, A_3 \oplus A_4)$	$\frac{7}{12}F$
$(1, 2)$	$(B_2, A_4 \oplus B_3)$	$\frac{7}{12}F$
$(1, 3)$	(A_4, C_2)	$\frac{7}{12}F$
$(2, 1)$	$(A_2, A_3 \oplus B_4)$	$\frac{7}{12}F$
$(2, 2)$	$(B_2, B_3 \oplus B_4)$	$\frac{7}{12}F$
$(2, 3)$	(B_4, C_2)	$\frac{7}{12}F$

Corollary 1 shows that in this particular scenario, an optimal unselfish design further reduces the delivery rate by 14% when compared to an optimal selfish solution.

C. Other Miscellaneous Results

For general Θ_k, exact capacity characterization of $\tilde{R}$ remains an open problem. In the following, we provide some partial results that do not have matching converse and achievable rates.

Proposition 6 (Converse w. $\Theta_1 \subsetneq \Theta_2$). *Consider $K = 2$ users and $\Theta_1 = \{1, \cdots, N_1\}$ and $\Theta_2 = \{1, \cdots, N_2\}$ satisfying*

$$3 \le 1.5N_1 \le N_2 \le 2N_1 \text{ and } N_1 \text{ and } N_2 \text{ being even.} \tag{8}$$

The average-rate $\tilde{R}$ of an unselfish scheme must satisfy

$$\tilde{R} \ge F - (M_1/N_1) \tag{P1+}$$
$$\tilde{R} \ge F - (M_2/N_2) \tag{P2+}$$
$$\tilde{R} \ge \frac{N_1 + N_2}{2N_1}F - \frac{M_1 + M_2}{2N_1} \tag{P3+}$$
$$\tilde{R} \ge \frac{3}{2}F - \frac{M_1 + M_2}{N_2} \tag{P4+}$$
$$\tilde{R} \ge \frac{N_1 + N_2}{N_2}F - \frac{M_1}{N_1} - \frac{M_2}{N_2} \tag{P5+}$$
$$\tilde{R} \ge \frac{N_1 + N_2}{N_2}F - \frac{M_1}{N_2} - \frac{3M_2}{2N_2} \tag{P6+}$$

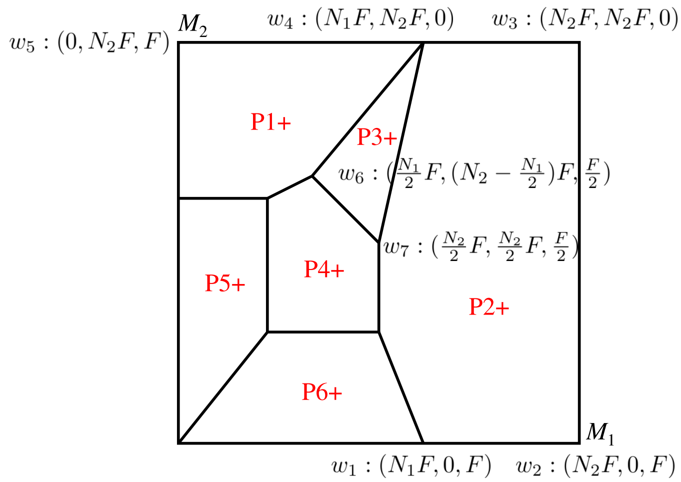

Fig. 5. The converse rate lower bounds of (P1+) to (P6+) w. $\Theta_1 = \{1, \cdots, N_1\}$ and $\Theta_2 = \{1, \cdots, N_2\}$ that satisfy (8). There are 12 vertices and we only label the 7 vertices for which we have a matching achievable rate.

Fig. 5 illustrates the converse (P1+) to (P6+).

Proposition 7 (Achievability w. $\Theta_1 \subsetneq \Theta_2$). *Continue from Proposition 6. The lower bounds form 12 vertices in Fig. 5. Among them, 7 vertices w_1 to w_7 are achievable.*

The proofs of Proposition 6 and 7 is delegated in Appendix F.

Comparing Proposition 6 and 7, we have characterized the exact $\tilde{R}$ for the 3 subregions, each surrounded by the vertex sets $\{w_4, w_5, w_6\}$, $\{w_4, w_6, w_7\}$, and $\{w_1, w_2, w_3, w_4, w_7\}$, respectively. I.e., when either M_1 or M_2 is sufficiently large. Note that when $(N_1, N_2) = (2, 3)$, then (P1+) to (P6+) collapse to (P1) to (P6) in Proposition 4, and they thus tightly characterize the capacity for all (M_1, M_2).

IV. Conclusion

We consider coded caching with heterogeneous file popularity. Each user k only desires the files in his/her file demand set (FDS) Θ_k with equal probability $\frac{1}{|\Theta_k|}$ and we investigate the average rate capacity in various scenarios. A byproduct is the first proof showing that in some simple setting, unselfish designs can strictly outperform selfish ones by 14%.

Appendix A
Proof of Proposition 1

We first prove the selfish and unselfish designs achieves the same average rate $\tilde{R}$. We construct the lower bounds following the cut-set bounds in [1] for disjoint Θ_k, $k \in [K]$. That is, for all $k \in [K]$, $S_k \subseteq \Theta_k$, and $T = \prod_{k \in K} |S_k|$, we have

$$\sum_{\vec{d}: d_k \in S_k} R_{\vec{d}} + \sum_{k \in [K]} \frac{T}{|S_k|} M_k \ge \sum_{k \in [K]} \frac{T}{|S_k|} \left(\sum_{i \in S_k} F_i \right). \tag{9}$$

If we let $F_i = F$ for all $i \in [N]$ and substitute the uniform average rate (7) in Definition 5 in (9), we obtain

$$\tilde{R} \geq \sum_{k\in[K]} \left(F - \frac{M_k}{|\Theta_k|}\right)^+. \tag{10}$$

The achievable scheme is as follows. In the placement phase, each user k caches $M_k/|\Theta_k|$ size of each file W_i, $i \in \Theta_k$, and in delivery phase, for any demand $\vec{d}$, the server transmits the remain $(F - \frac{M_k}{|\Theta_k|})^+$ fraction of file for each demanded file W_{d_k} with rate $R_{\vec{d}} = \sum_{k\in[K]}(F - \frac{M_k}{|\Theta_k|})^+$ and hence the uniform-average rate is achieved with equality in (10).

For the worst-case rate R^*, it is clear that $R^* \geq \tilde{R}$ such that we also have

$$R^* \geq \sum_{k\in[K]} \left(F - \frac{M_k}{|\Theta_k|}\right)^+. \tag{11}$$

Since the described achievable scheme has same $R_{\vec{d}} = \sum_{k\in[K]}(F - \frac{M_k}{|\Theta_k|})^+$ for all $\vec{d}$, the worst-case rate is achieved with equality in (11).

Appendix B
Proof of Proposition 2

We prove the uniform-average rate capacity R^* by first deriving the rate lower bounds (Q1) to (Q4) and then showing that all the corner points formed by the bounds are achievable. The bounds (Q1) and (Q2) are the cut-set bounds in [1]–[3]. Therefore we focus on the bound (Q3) (and the symmetrical form (Q4)).

To prove (Q3), we first consider the following inequality

$$\begin{aligned}
&\sum_{j\in[N]\setminus\{i\}} H(X_{(i,j)}, Z_1) = \sum_{j\in[N]\setminus\{i\}} H(X_{(i,j)}, Z_1, W_i) &(12)\\
&\geq (N-2)H(Z_1, W_i) + H([X_{(i,j)}]_{j\in[N]\setminus\{i\}}, Z_1, W_i) &(13)\\
&\geq (N-2)H(Z_1, W_i) + H([X_{(i,j)}]_{j\in[N]\setminus\{i\}}, Z_1, Z_2, W_i)\\
&\quad + H(W_i) - H(Z_2, W_i) &(14)\\
&\geq (N-2)H(Z_1, W_i) + H([W_j]_{j\in[N]})\\
&\quad + H(W_i) - H(Z_2, W_i) &(15)\\
&= (N+1)F + (N-2)H(Z_1, W_i) - H(Z_2, W_i) &(16)
\end{aligned}$$

where (12) follows from that user 1 can decode W_i based on $X_{(i,j)}$ and Z_1; (13) and (14) follows from using the matroidal Shannon inequality. (15) follows from that user 2 can decode $[W_j]_{j\in[N]\setminus\{i\}}$ based on $[X_{(i,j)}]_{j\in[N]\setminus\{i\}}$ and Z_2.

Following (16), we obtain

$$\begin{aligned}
&N(N-1)M_1 + \sum_{i\in[N]}\sum_{j\in[N]\setminus\{i\}} R_{(i,j)}\\
&\geq \sum_{i\in[N]}\sum_{j\in[N]\setminus\{i\}} H(X_{(i,j)}, Z_1) \geq N(N+1)F\\
&\quad + (N-2)\sum_{i\in[N]} H(Z_1, W_i) - \sum_{i\in[N]} H(Z_2, W_i), &(17)
\end{aligned}$$

and symmetrically we also have

$$\begin{aligned}
&N(N-1)M_2 + \sum_{i\in[N]}\sum_{j\in[N]\setminus\{i\}} R_{(i,j)} \geq N(N+1)F\\
&\quad + (N-2)\sum_{i\in[N]} H(Z_2, W_i) - \sum_{i\in[N]} H(Z_1, W_i). &(18)
\end{aligned}$$

Computing $((17) + (N-2)\times(18))/(N-1)$ yields

$$\begin{aligned}
&NM_1 + N(N-2)M_2 + \sum_{i\in[N]}\sum_{j\in[N]\setminus\{i\}} R_{(i,j)} &(19)\\
&\geq N(N+1)F + (N-3)\sum_{i\in[N]} H(Z_2, W_i) &(20)\\
&\geq N(N+1)F\\
&\quad + (N-3)\big((N-1)H(Z_2) + H(Z_2, [W_i]_{i\in[N]})\big) &(21)\\
&\geq 2N(N-1)F + (N-3)(N-1)M_2 &(22)
\end{aligned}$$

or equivalently

$$NM_1 + (2N-3)M_2 + \sum_{i,j\in[N], i\neq j} R_{(i,j)} \geq 2N(N-1)F \tag{23}$$

Adding (23) with the cut-set bound

$$M_2 + \sum_{i\in[N]} R_{(i,i)} \geq NF \tag{24}$$

leads to

$$NM_1 + 2(N-1)M_2 + \sum_{i,j\in[N]} R_{(i,j)} \geq N(2N-1)F. \tag{25}$$

By Definition 5, we substitute $\tilde{R} = \frac{1}{N^2}\sum_{i,j\in[N]} R_{(i,j)}$ in to (25) to obtain (Q3).

One can verify the lower bounds (Q1) to (Q4) with respect to $M_1, M_2 \in [0, NF]$ is the surface plotted in Fig. 1. Now we show that joint (Q1) to (Q4) is the capacity by verifying all the corner points t_1 to t_7 in Fig. 1 are zero-error achievable.

The achievable scheme for point t_1 is trivial. The points t_2 and t_3 can be zero-error achieved by user 1 caches $Z_1 = (W_1 \oplus W_2, W_1 \oplus W_3, \cdots, W_1 \oplus W_N)$ of size $(N-1)F$ in the placement phase and in the delivery phase, for the demand $(d_1, d_2) \in [N]^2$, the server transmits $X_{(d_1,d_2)} = W_{d_2}$ corresponding to rate $R_{(d_1,d_2)} = F$ and hence $\tilde{R} = F$. The points t_4 can be zero-error achieved by user 1 and 2 both caches $Z_1 = Z_2 = (W_1, W_2, \ldots, W_N)$ of size NF in the placement phase and for any demand patterns, no need for extra transmission in the delivery phase. The achievable scheme of t_5 and t_6 are the user symmetric scheme of t_3 and t_2, respectively. The point t_7 can be zero-error achieved by first divide the N files into two halves: $W_1 = (U_1, V_1), \ldots, W_N = (U_N, V_N)$. In the placement phase, user 1 caches $Z_1 = (U_1, U_2, \ldots, U_N)$ and user 2 caches $Z_2 = (V_1, V_2, \ldots, V_N)$; and in the delivery phase for the demand $(d_1, d_2) \in [N]^2$, the server transmits $X_{(d_1,d_2)} = U_{d_2} \oplus V_{d_1}$ corresponding to rate $R_{(d_1,d_2)} = \frac{F}{2}$ and hence $\tilde{R} = \frac{F}{2}$.

Since $\Theta_1 = \Theta_2 = [N]$, it is clear that all the achievable schemes are selfish, i.e., the uniform average rate capacities for selfish and unselfish designs are the same.

Appendix C
Proof of Proposition 3

We prove Proposition 3 by showing that (Q1) to (Q3) are the lower bounds of $\tilde{R}$, and then showing all the rates satisfying (Q1) to (Q3) are zero-error achievable.

We first derive Q_1 and Q_2 from the cut-set bounds. For $i \in [N]$, we have the cut-set bounds

$$M_1 + R_{(1,i)} \geq H(X_{(1,i)}, Z_1) \geq H(W_1) = F \tag{26}$$

and hence the bound Q_1 is obtained by the summation $NM_1 + N\tilde{R} = NM_1 + \sum_{i=1}^{N} R_{(1,i)} \geq NF$. The bound Q_2 is deduced by another cut-set bound

$$M_2 + \sum_{i=1}^{N} R_{(1,i)} \geq H([X_{(1,i)}]_{i=1}^{N}, Z_2) \geq H([W_i]_{i=1}^{N}) = NF$$

and hence we have the Q_2 bound $M_2 + N\tilde{R} \geq NF$.

The bounds (Q1) and (Q2) are the cut set bounds following the close argument in [1]. The bound (Q3) can be derived by

$$\begin{aligned}
&(N-1)M_1 + M_2 + N\tilde{R} && (27)\\
&\geq \sum_{i=2}^{N} H(Z_1, X_{(1,i)}) + H(Z_2, X_{(1,1)}) && (28)\\
&= \sum_{i=2}^{N} H(Z_1, X_{(1,i)}, W_1) + H(Z_2, X_{(1,1)}, W_1) && (29)\\
&\geq (N-2)H(Z_1, W_1) + H(Z_1, [X_{(1,i)}]_{i=2}^{N}, W_1)\\
&\quad + H(Z_2, X_{(1,1)}, W_1) && (30)\\
&\geq (N-1)H(W_1) + H(Z_1, Z_2, [X_{(1,i)}]_{i=1}^{N}, W_1) && (31)\\
&= (N-1)H(W_1) + H(Z_1, Z_2, [X_{(1,i)}]_{i=1}^{N}, [W_i]_{i=1}^{N}) && (32)\\
&\geq (N-1)H(W_1) + H([W_i]_{i=1}^{N}) = (2N-1)F && (33)
\end{aligned}$$

where (28) follows from the definition of average rate $\tilde{R} = \frac{1}{N}\sum_{i=1}^{N} H(X_{(1,i)})$; (29) follows from that user 1 can decode W_1 based on $X_{(1,i)}$ and Z_1 and user 2 can decode W_1 based on $X_{(1,1)}$ and Z_2. (30) follows from using the matroidal Shannon inequality $N-2$ times on the term $\sum_{i=2}^{N} H(Z_1, X_{(1,i)}, W_1)$; (31) follows from the inequalities $H(Z_1, W_1) \geq H(W_1)$ and the matroidal Shannon inequality on $H(Z_1, [X_{(1,i)}]_{i=2}^{N}, W_1)$ and $H(Z_2, X_{(1,1)}, W_1)$. (32) follows from that user 2 can decode W_i based on $X_{(1,i)}$. Therefore we prove (Q3) is a lower bound of average rate

One can verify the lower bounds (Q1) to (Q3) with respect to $M_1, M_2 \in [0, NF]$ is the surface as plotted in Fig. 2. Now we show that joint (Q1) to (Q3) is the capacity by verifying all the corner points u_1 to u_7 in Fig. 2 are zero-error achievable.

The achievable scheme for point u_1 is trivial. The points u_2 and u_3 can be zero-error achieved by user 1 caches $Z_1 = W_1$ of size F in the placement phase and in the delivery phase, for the demand $(d_1, d_2) = (1, i)$, $i \in [N]$, the server transmits $X_{(1,i)} = W_i$ corresponding to rate $R_{(1,i)} = F$. The points u_4 and u_5 can be zero-error achieved by user 1 caches $Z_1 = W_1$ of size F and user 2 caches $Z_2 = (W_1, W_2, \ldots, W_N)$ of size NF in the placement phase and for any demand patterns, no need for extra transmission in the delivery phase. The points u_6 and u_7 can be zero-error achieved by user 2 caches $Z_2 = (W_2, \ldots, W_N)$ of size $(N-1)F$ in the placement phase and in the delivery phase, for the demand $(d_1, d_2) = (1, i)$, $i \in [N]$, the server transmits $X_{(1,i)} = W_1$ corresponding to rate $R_{(1,i)} = F$.

It is clear that the proposed achievable schemes of corner points u_1 to u_7 are actually selfish coded caching schemes. Therefore the uniform average rate capacities for selfish and unselfish designs are the same.

Appendix D
Proof of Proposition 4

The bounds (P1) and (P2) are the cut set bounds following the close argument in [1]. To derive bound (P3), we first obtain following inequality for $\{3\} \subseteq \{i, j\} \subseteq \{1, 2, 3\}$ and $k, l \in \{1, 2\}$.

$$\begin{aligned}
&M_1 + M_2 + R_{(1,i)} + R_{(2,j)} + R_{(k,1)} + R_{(l,2)} && (34)\\
&\geq H(X_{(1,i)}, X_{(2,j)}, Z_1) + H(X_{(k,1)}, X_{(l,2)}, Z_2) && (35)\\
&= H(X_{(1,i)}, X_{(2,j)}, Z_1, W_1, W_2)\\
&\quad + H(X_{(k,1)}, X_{(l,2)}, Z_2, W_1, W_2) && (36)\\
&\geq H(X_{(1,i)}, X_{(2,j)}, X_{(k,1)}, X_{(l,2)}, Z_2, W_1, W_2)\\
&\quad + H(W_1, W_2) && (37)\\
&= H(X_{(1,i)}, X_{(2,j)}, X_{(k,1)}, X_{(l,2)}, Z_2, W_1, W_2, W_3)\\
&\quad + H(W_1, W_2) && (38)\\
&\geq H(W_1, W_2, W_3) + H(W_1, W_2) = 5F && (39)
\end{aligned}$$

where (36) follows from that user 1 can decode W_1 and W_2 based on $X_{(1,i)}$, $X_{(2,j)}$, and Z_1; and user 2 can decode W_1 and W_2 based on $X_{(k,1)}$, $X_{(l,2)}$, and Z_2. (37) follows from using the matroidal Shannon inequality. (38) follows from the assumption $i = 3$ or $j = 3$ such that user 2 can decode W_3 based on $X_{(1,i)}$, $X_{(2,j)}$, and Z_2.

We then obtain three inequalities for $(i, j, k, l) = (3, 1, 1, 2)$, $(2, 3, 1, 2)$, and $(3, 3, 2, 1)$, respectively as follows.

$$M_1 + M_2 + R_{(1,3)} + R_{(2,1)} + R_{(1,1)} + R_{(2,2)} \geq 5F \tag{40}$$
$$M_1 + M_2 + R_{(1,2)} + R_{(2,3)} + R_{(1,1)} + R_{(2,2)} \geq 5F \tag{41}$$
$$M_1 + M_2 + R_{(1,3)} + R_{(2,3)} + R_{(2,1)} + R_{(1,2)} \geq 5F. \tag{42}$$

The summation of (40), (41), and (42) yields $3M_1 + 3M_2 + 12\tilde{R} \geq 15F$ or equivalently (P3).

To derive the bound (P4), we first obtain the following inequality for $(i, j) = (1, 2)$ or $(2, 1)$.

$$\begin{aligned}
&M_1 + M_2 + R_{(j,3)} + R_{(i,j)} && (43)\\
&\geq H(X_{(j,3)}, Z_1) + H(X_{(i,j)}, Z_2) && (44)\\
&= H(X_{(j,3)}, Z_1, W_j) + H(X_{(i,j)}, Z_2, W_j) && (45)\\
&\geq H(X_{(j,3)}, X_{(i,j)}, Z_1, Z_2, W_j) + H(W_j) && (46)\\
&= H(X_{(j,3)}, X_{(i,j)}, Z_1, Z_2, W_i, W_j, W_3) + H(W_j) && (47)\\
&\geq H(W_1, W_2, W_3) + H(W_j) = 4F. && (48)
\end{aligned}$$

where (45) follows from that user 1 can decode W_j based on $X_{(j,3)}$ and Z_1; and user 2 can decode W_j based on $X_{(i,j)}$ and Z_2. (46) follows from using the matroidal Shannon inequality. (47) follows from the assumption $(i, j) = (1, 2)$ or $(2, 1)$ such that user 1 can decode W_i based on $X_{(i,j)}$ and Z_1; and user 2 can decode W_3 based on $X_{(j,3)}$ and Z_2.

We then obtain two inequalities for $(i, j) = (1, 2)$ and $(2, 1)$, respectively as follows.

$$M_1 + M_2 + R_{(2,3)} + R_{(1,2)} \geq 4F \tag{49}$$
$$M_1 + M_2 + R_{(1,3)} + R_{(2,1)} \geq 4F. \tag{50}$$

The summation of (40) and (49) yields $2M_1+2M_2+6\tilde{R} \geq 9F$ or equivalently (P4).

To derive the bounds (P5) and (P6), we first obtain the following cut-set bounds

$$M_1 + R_{(1,1)} + R_{(2,2)} \geq 2F \tag{51}$$
$$M_2 + R_{(1,1)} + R_{(2,2)} \geq 2F. \tag{52}$$

The summation of (49), (50), and (96) yields (P5) and the summation of (49), (50), and (52) yields (P6).

Now we show that corner points v_1 to v_{11} of bounds (P1) to (P6) are zero-error achievable. The achievable scheme for point v_1 is trivial. The points v_2 and v_3 can be zero-error achieved by user 1 caches $Z_1 = (W_1, W_2)$ of size $2F$ in the placement phase and in the delivery phase, for any demand $(d_1, d_2) = (i, j)$, the server transmits $X_{(i,j)} = W_j$ corresponding to rate $R_{(i,j)} = F$. The points v_4 and v_5 can be zero-error achieved by user 1 caches $Z_1 = (W_1, W_2)$ of size $2F$ and user 2 caches $Z_2 = (W_1, W_2, W_3)$ of size $3F$ in the placement phase and for any demand patterns, no need for extra transmission in the delivery phase. The points v_6 and v_7 can be zero-error achieved by user 2 caches $Z_2 = (W_1 \oplus W_2, W_1 \oplus W_3)$ of size $2F$ in the placement phase and in the delivery phase, for the demand $(d_1, d_2) = (i, j)$, the server transmits $X_{(i,j)} = W_i$ corresponding to rate $R_{(i,j)} = F$.

For describing the achievable schemes of corner points v_8 to v_{11}, we divide each files into two disjoint subfiles of equal size $\frac{F}{2}$, i.e., $W_1 = (A_1, A_2)$, $W_2 = (B_1, B_2)$, and $W_3 = (C_1, C_2)$. The zero-error achievable scheme of v_8 is that in the placement phase, user 1 caches $Z_1 = (A_1, B_1)$ of size F, and user 2 caches $Z_2 = (A_2, B_2, W_3)$ of size $2F$ and in the delivery phase, for the 6 possible demands the server transmits $X_{(1,1)} = A_1 \oplus A_2$, $X_{(1,2)} = A_2 \oplus B_1$, $X_{(1,3)} = A_2$, $X_{(2,1)} = A_1 \oplus B_2$, $X_{(2,2)} = B_1 \oplus B_2$, and $X_{(2,3)} = B_2$ corresponding to the same rate $\frac{F}{2}$ and hence achieves the average rate $\tilde{R} = \frac{F}{2}$. The zero-error achievable scheme of v_9 is that in the placement phase, user 1 caches $Z_1 = (A_1, B_1)$ of size F, and user 2 caches $Z_2 = (A_2, B_2)$ of size F and in the delivery phase, for the 6 possible demands the server transmits $X_{(1,1)} = A_1 \oplus A_2$, $X_{(1,2)} = A_2 \oplus B_1$, $X_{(1,3)} = (A_2, W_3)$, $X_{(2,1)} = A_1 \oplus B_2$, $X_{(2,2)} = B_1 \oplus B_2$, and $X_{(2,3)} = (B_2, W_3)$ corresponding to the rates $R_{(1,1)} = R_{(1,2)} = R_{(2,1)} = R_{(2,2)} = \frac{F}{2}$, and $R_{(1,3)} = R_{(2,3)} = \frac{3F}{2}$. The average rate is therefore $\tilde{R} = \frac{5}{6}F$.

The zero-error achievable scheme of v_{10} is that in the placement phase, user 1 caches $Z_1 = (A_1, B_1, C_2 \oplus A_2 \oplus B_2)$ of size $\frac{3F}{2}$, and user 2 caches $Z_2 = (A_2, B_2)$ of size F and in the delivery phase, for the 6 possible demands the server transmits $X_{(1,1)} = A_1 \oplus A_2$, $X_{(1,2)} = A_2 \oplus B_1$, $X_{(1,3)} = (C_1, C_2 \oplus B_2)$, $X_{(2,1)} = A_1 \oplus B_2$, $X_{(2,2)} = B_1 \oplus B_2$, and $X_{(2,3)} = (C_1, C_2 \oplus A_2)$ corresponding to the rates $R_{(1,1)} = R_{(1,2)} = R_{(2,1)} = R_{(2,2)} = \frac{F}{2}$, and $R_{(1,3)} = R_{(2,3)} = F$. The average rate is therefore $\tilde{R} = \frac{2}{3}F$. The zero-error achievable scheme of v_{11} is that in the placement phase, user 1 caches $Z_1 = (A_1, B_1, C_1)$ of size $\frac{3}{2}F$, and user 2 caches $Z_2 = (A_2, B_2, C_2)$ of size $\frac{3}{2}F$ and in the delivery phase, for the 6 possible demands the server transmits $X_{(1,1)} = A_1 \oplus A_2$, $X_{(1,2)} = A_2 \oplus B_1$, $X_{(1,3)} = A_2 \oplus C_1$, $X_{(2,1)} = A_1 \oplus B_2$, $X_{(2,2)} = B_1 \oplus B_2$, and $X_{(2,3)} = B_2 \oplus C_1$ corresponding to the same rate $\frac{F}{2}$ and hence achieves the average rate $\tilde{R} = \frac{F}{2}$.

Appendix E
Proof of Proposition 5

The bound (P7), we first obtain the following inequality for $i \in \{1, 2\}$

$$\begin{aligned}
&M_1 + M_2 + R_{(i,1)} + R_{(i,2)} + R_{(i,3)} && (53)\\
&\geq H(X_{(i,3)}, Z_1) + H(X_{(i,1)}, X_{(i,2)}, Z_2) && (54)\\
&= H(X_{(i,3)}, Z_1, W_i) + H(X_{(i,1)}, X_{(i,2)}, Z_2, W_1, W_2) && (55)\\
&= H(X_{(i,3)}, Z_1, W_i) \\
&\quad + H(X_{(i,1)}, X_{(i,2)}, Z_1, Z_2, W_1, W_2) && (56)\\
&\geq H(X_{(i,3)}, X_{(i,1)}, X_{(i,2)}, Z_1, Z_2, W_1, W_2) \\
&\quad + H(Z_1, W_i) && (57)\\
&= H(X_{(i,3)}, X_{(i,1)}, X_{(i,2)}, Z_1, Z_2, W_1, W_2, W_3) \\
&\quad + H(Z_1, W_i) && (58)\\
&\geq H(W_1, W_2, W_3) + H(Z_1, W_i) = 3F + H(Z_1, W_i) && (59)
\end{aligned}$$

where (55) follows from that user 1 can decode W_i based on $X_{(i,3)}$ and Z_1; and user 2 can decode W_1 and W_2 based on $X_{(i,1)}$, $X_{(i,2)}$, and Z_2. (56) follows from the definition of selfish coded caching where $Z_1 = \phi_1(W_1, W_2)$. (57) follows from the assumption $i \in \{1, 2\}$ and the matroidal Shannon inequality. (58) follows from user 2 can decode W_3 based on $X_{(i,3)}$ and Z_2.

We then obtain two inequalities for $i = 1$ and $i = 2$ as follows

$$M_1 + M_2 + \sum_{j=1}^{3} R_{(1,j)} \geq 3F + H(Z_1, W_1) \tag{60}$$

$$M_1 + M_2 + \sum_{j=1}^{3} R_{(2,j)} \geq 3F + H(Z_1, W_2). \tag{61}$$

The summation of (60) and (61) leads to

$$\begin{aligned}
&2M_1 + 2M_2 + 6\tilde{R} && (62)\\
&\geq 6F + H(Z_1, W_1) + H(Z_1, W_2) && (63)\\
&\geq 6F + H(Z_1, W_1, W_2) + H(Z_1) && (64)\\
&\geq 6F + H(W_1, W_2) + H(Z_1) = 8F + M_1 && (65)
\end{aligned}$$

and hence the bound (P7), where (64) follows from the matroidal Shannon inequality.

It is clear that the aforementioned achievable schemes of corner points v_1 to v_7 are actually selfish coded caching schemes. Therefore joint (P1), (P2), and (P5) to (P7) is the selfish capacity of uniform average rate.

Appendix F
Proof of Proposition 6 and 7

We prove Proposition 6 by showing that (P1+) to (P6+) are the rate lower bounds of unselfish coded caching schemes. The bounds (P1+) and (P2+) can derived from the close argument in [1].

Given the condition $N_1 \leq 2N_1$, for any set $\{j_i \in [N_2] : i \in [N_1]\} \supseteq [N_2]\backslash[N_1]$ and $\{k_i \in [N_1] : i \in [N_1]\}$, the bound (P3+) can be proved by

$$\begin{aligned} & M_1 + M_2 + \sum_{i\in[N_1], j_i, k_i \in [N_2]} \left(R_{(i,j_i)} + R_{(k_i,i)}\right) && (66)\\ & \geq H([X_{(i,j_i)}]_{i\in[N_1]}, Z_1) + H([X_{(k_i,i)}]_{i\in[N_1]}, Z_2) && (67)\\ & = H([X_{(i,j_i)}]_{i\in[N_1]}, Z_1, [W_i]_{i\in[N_1]}) \\ & \quad + H([X_{(k_i,i)}]_{i\in[N_1]}, Z_2, [W_i]_{i\in[N_1]}) && (68)\\ & \geq H([X_{(i,j_i)}]_{i\in[N_1]}, [X_{(k_i,i)}]_{i\in[N_1]}, Z_1, Z_2, [W_i]_{i\in[N_1]}) \\ & \quad + H([W_i]_{i\in[N_1]}) && (69)\\ & = H([X_{(i,j_i)}]_{i\in[N_1]}, [X_{(k_i,i)}]_{i\in[N_1]}, Z_1, Z_2, [W_i]_{i\in[N_2]}) \\ & \quad + H([W_i]_{i\in[N_1]}) && (70)\\ & = H([W_i]_{i\in[N_2]}) + H([W_i]_{i\in[N_1]}) = (N_1+N_2)F && (71) \end{aligned}$$

where (68) follows from that user 1 can decode $\{W_1, \ldots, W_{N_1}\}$ based on $\{[X_{(i,j_i)}]_{i\in[N_1]}, Z_1\}$ and user 2 can decode $\{W_1, \ldots, W_{N_1}\}$ based on $\{[X_{(k_i,i)}]_{i\in[N_1]}, Z_2\}$; (69) follows from the matroidal Shannon inequality; (70) follows that user 2 can decode $\{W_{N_1+1}, \ldots, W_{N_2}\}$ based on $\{[X_{(i,j_i)}]_{i\in[N_1]}, [X_{(k_i,i)}]_{i\in[N_1]}, Z_1\}$ due to $\{j_i\} \supseteq [N_2]\backslash[N_1]$.

We then show that following a procedure of randomly assignment of $\{j_i\}$ and $\{k_i\}$, we can uniformly cover all the demands $(d_1, d_2) \in [N_1] \times [N_2]$. To construct $R_{(i,j_i)}$ for fixed and ordered $i = 1, \ldots, N_1$, we first randomly pick $N_2 - N_1$ elements from $\{j_i\}$, and randomly one-to-one mapping to $\{N_1+1, \ldots, N_2\}$. Then remained $2N_1 - N_2 \geq 0$ elements in $\{j_i\}$ are unassigned. We then assign them to $2N_1 - N_2$ randomly-picked elements in $[N_2]$. To construct $R_{(k_i,i)}$ for fixed and ordered $i = 1, \ldots, N_1$, we randomly pick N_1 elements from $[N_1]$ and map to $\{k_i\}$. In this way, a demand (d_1, d_2), $d_1, d_2 \in [N_1]$, corresponding to the rate $R_{(d_1,d_2)}$ appeals with the probability

$$\frac{2N_1 - N_2}{N_1}\frac{1}{N_2} + \frac{1}{N_1} = \frac{2}{N_2} \tag{72}$$

and for $d_1 \in [N_1]$, $d_2 \in [N_2]\backslash[N_1]$, $R_{(d_1,d_2)}$ appeals with the probability

$$\frac{N_2 - N_1}{N_1}\frac{1}{N_2 - N_1} + \frac{2N_1 - N_2}{N_1}\frac{1}{N_2} = \frac{2}{N_2}. \tag{73}$$

Therefore if we forms sufficient large number of inequalities (71) with such random $\{j_i\}$ and $\{k_i\}$, we can replace $\sum_{i\in[N_1]} \left(R_{(i,j_i)} + R_{(k_i,i)}\right)$ in (66) with $\frac{2}{N_2}\sum_{i\in[N_1]}\sum_{j\in[N_2]} R_{(i,j)} = 2N_1\tilde{R}$ such that (71) becomes $M_1 + M_2 + 2N_1\tilde{R} \geq (N_1+N_2)F$ or equivalently (P3+).

Given the condition $N_2 \leq 2N_1$, we prove the bound (P4+) from the following inequality for a set $I \subseteq [N_1]$, $|I| = \lfloor N_2/2 \rfloor \leq N_1$, $\{j_i \in [N_2] : i \in I\}$, and $\{k_i \in [N_1] : i \in I\}$ such that $\{j_i\} \cup \{k_i\} \supseteq [N_2]\backslash I$

$$\begin{aligned} & M_1 + M_2 + \sum_{i\in I} \left(R_{(i,j_i)} + R_{(k_i,i)}\right) && (74)\\ & \geq H([X_{(i,j_i)}]_{i\in I}, Z_1) + H([X_{(k_i,i)}]_{i\in I}, Z_2) && (75)\\ & \geq H([X_{(i,j_i)}]_{i\in I}, Z_1, [W_i]_{i\in I}) \\ & \quad + H([X_{(k_i,i)}]_{i\in I}, Z_2, [W_i]_{i\in I}) && (76)\\ & \geq H([X_{(i,j_i)}]_{i\in I}, [X_{(k_i,i)}]_{i\in I}, Z_1, [W_i]_{i\in I}) \\ & \quad + H([W_i]_{i\in I}) && (77)\\ & \geq H([X_{(i,j_i)}]_{i\in I}, [X_{(k_i,i)}]_{i\in I}, Z_1, [W_i]_{i\in[N_2]}) \\ & \quad + H([W_i]_{i\in I}) && (78)\\ & = H([W_i]_{i\in[N_2]}) + H([W_i]_{i\in I}) = (N_2 + \lfloor N_2/2 \rfloor)F && (79) \end{aligned}$$

where the (76) follows from that user 1 can decode $\{W_i : i \in I\}$ based on $\{[X_{(i,j_i)}]_{i\in I}, Z_1\}$ and user 2 can decode $\{W_i : i \in I\}$ based on $\{[X_{(k_i,i)}]_{i\in I}, Z_2\}$; (77) follows from the matroidal Shannon inequality; (78) follows that user 1 and user 2 can decode $\{W_i : i \in [N_2]\backslash I\}$ based on $\{[X_{(i,j_i)}]_{i\in[N_1]}, [X_{(k_i,i)}]_{i\in[N_1]}, Z_1, Z_2\}$ due to $\{j_i\} \cup \{k_i\} \supseteq [N_2]\backslash I$.

We construct the set I, $\{j_i\}$, and $\{k_i\}$ as follows. We first randomly choose $\lfloor N_2/2 \rfloor - 1$ elements from $[N_1]$ without repetition and permute them to an ordered set I. We then set $k_{i_1} = i_1$ and $(k_{i_2}, \ldots, k_{i_{N_2-N_1+1}}) = (N_1+1, \ldots, N_2)$. We then randomly pick $N_1 + 2\lfloor N_2/2 \rfloor - N_2 - 1$ elements with repetition in $[N_1]\backslash I$ to assign the rest of $\{k_i\} \cup \{j_i\}$ such that $\{k_i\} \cup \{j_i\} \supseteq [N_1]\backslash I$. In this way, a demand (d_1, d_2), $d_1 \in [N_1]$, $d_2 \in [N_2]$ corresponding to the rate $R_{(d_1,d_2)}$ appeals with the probability

$$\begin{cases} \frac{1}{N_1} & d_1 = d_2 \in [N_1] \\ \frac{1}{N_1}\frac{N_1 - \lfloor N_2/2 \rfloor}{N_1 - 1}\frac{N_1 + 2\lfloor N_2/2 \rfloor - N_2 - 1}{N_1 - \lfloor N_2/2 \rfloor} & d_1 \neq d_2 \in [N_1] \\ \frac{1}{N_1} & d_1 \in [N_1], d_2 \in [N_2]. \end{cases} \tag{80}$$

On the other hand, we can obtain another inequality for a set $L \subseteq [N_1]$, $|L| = \lceil N_2/2 \rceil \leq N_1$, $\{j_l \in [N_2] : l \in L\} \supseteq [N_2]$, and $\{k_l \in [N_1] : l \in L\}$ such that $\{j_l\} \cup \{k_l\} \supseteq [N_2]\backslash L$

$$\begin{aligned} & M_1 + M_2 + \sum_{i\in L} \left(R_{(l,j_l)} + R_{(k_l,l)}\right) && (81)\\ & \geq H([X_{(l,j_l)}]_{l\in L}, Z_1) + H([X_{(k_l,l)}]_{l\in L}, Z_2) && (82)\\ & \geq H([X_{(l,j_l)}]_{l\in L}, Z_1, [W_l]_{l\in L}) \\ & \quad + H([X_{(k_l,l)}]_{l\in L}, Z_2, [W_l]_{l\in L}) && (83)\\ & \geq H([X_{(l,j_l)}]_{l\in L}, [X_{(k_l,l)}]_{l\in L}, Z_1, [W_l]_{l\in L}) \\ & \quad + H([W_l]_{l\in L}) && (84)\\ & \geq H([X_{(l,j_l)}]_{l\in L}, [X_{(k_l,l)}]_{l\in L}, Z_1, [W_l]_{l\in[N_2]}) \\ & \quad + H([W_l]_{l\in L}) && (85)\\ & = H([W_l]_{l\in[N_2]}) + H([W_l]_{l\in L}) = (N_2 + \lceil N_2/2 \rceil)F && (86) \end{aligned}$$

where the (83) follows from that user 1 can decode $\{W_l : l \in L\}$ based on $\{[X_{(l,j_l)}]_{l\in L}, Z_1\}$ and user 2 can decode $\{W_l : l \in L\}$ based on $\{[X_{(k_l,l)}]_{l\in L}, Z_2\}$; (84) follows from the matroidal Shannon inequality; (85) follows that user 2 can decode $\{W_l : l \in [N_2]\backslash L\}$ based on $\{[X_{(l,j_l)}]_{l\in[N_1]}, [X_{(k_l,l)}]_{l\in[N_1]}, Z_2\}$ due to $\{j_l\} \cup \{k_l\} \supseteq [N_2]\backslash L$.

We follow similar procedure to construct the set L, $\{j_l\}$, and $\{k_l\}$ as follows. We first randomly choose $\lceil N_2/2\rceil$ elements from $[N_1]$ without repetition and permute them to an ordered set L. We then set $k_{l_1}=l_1$ and $(k_{l_2},\ldots,k_{l_{N_2-N_1+1}})=(N_1+1,\ldots,N_2)$. We then randomly pick $N_1+2\lceil N_2/2\rceil-N_2-1$ elements with repetition in $[N_1]\backslash I$ to assign the rest of $\{k_i\}\cup\{j_i\}$ such that $\{k_i\}\cup\{j_i\}\supseteq[N_1]\backslash I$. In this way, a demand (d_1,d_2), $d_1\in[N_1]$, $d_2\in[N_2]$ corresponding to the rate $R_{(d_1,d_2)}$ appeals with the probability

$$\begin{cases}\frac{1}{N_1} & d_1=d_2\in[N_1]\\ \frac{1}{N_1}\frac{N_1-\lceil N_2/2\rceil}{N_1-1}\frac{N_1+2\lceil N_2/2\rceil-N_2-1}{N_1-\lceil N_2/2\rceil} & d_1\neq d_2\in[N_1]\\ \frac{1}{N_1} & d_1\in[N_1],d_2\in[N_2]\end{cases} \tag{87}$$

Add the inequality (79) and (86) together yields

$$\begin{aligned}&2M_1+2M_2+\sum_{i\in I}R_{(i,j_i)}+R_{(k_i,i)}+\sum_{i\in L}R_{(l,j_l)}+R_{(k_l,l)}\\&\geq(2N_2+\lfloor N_2/2\rfloor+\lceil N_2/2\rceil)F=3N_2F\end{aligned} \tag{88}$$

where a demand (d_1,d_2), $d_1\in[N_1]$, $d_2\in[N_2]$ corresponding to the rate $R_{(d_1,d_2)}$ appeals with probability sum of (80) and (87)

$$\begin{cases}\frac{1}{N_1}+\frac{1}{N_1}=\frac{2}{N_1} & d_1=d_2\in[N_1]\\ \frac{1}{N_1}\frac{N_1+2\lceil N_2/2\rceil-N_2-1}{N_1-1}+\frac{1}{N_1}\frac{N_1+2\lceil N_2/2\rceil-N_2-1}{N_1-1}=\frac{2}{N_1} & \\ & d_1\neq d_2\in[N_1]\\ \frac{1}{N_1}+\frac{1}{N_1}=\frac{2}{N_1} & d_1\in[N_1],d_2\in[N_2].\end{cases} \tag{89}$$

That is, for all (d_1,d_2), $d_1\in[N_1]$, $d_2\in[N_2]$, the rate $R_{(d_1,d_2)}$ appeals in (88) with probability $\frac{2}{N_1}$ such that we can replace $\sum_{i\in I}R_{(i,j_i)}+R_{(k_i,i)}+\sum_{i\in L}R_{(l,j_l)}+R_{(k_l,l)}$ with $\frac{2}{N_1}\sum_{i\in[N_1]}\sum_{j\in[N_2]}R_{(i,j)}=2N_2\tilde{R}$ in (88) and hence we obtain (P4+).

To prove (P5+), we consider even N_1 and $N_2>\frac{3}{2}N_1$. Let set $I\subset[N_1]$, $|I|=\frac{N_1}{2}$ be a subset of $[N_1]$ of $N_1/2$ elements, we denote sets $\{j_i:i\in I\}=[N_1]\backslash I$ and $\{k_i\in[N_2]\backslash[N_1]:i\in I\}$ both of distinct $N_1/2$ elements. We first derive the following inequality.

$$M_1+M_2+\sum_{i\in I}R_{(i,k_i)}+\sum_{i\in I}R_{(j_i,i)} \tag{90}$$
$$\geq H([X_{(i,k_i)}]_{i\in I},Z_1)+H([X_{(j_i,i)}]_{i\in I},Z_2) \tag{91}$$
$$\begin{aligned}&=H([X_{(i,k_i)}]_{i\in I},Z_1,[W_i]_{i\in I})\\&\quad+H([X_{(j_i,i)}]_{i\in I},Z_2,[W_i]_{i\in I})\end{aligned} \tag{92}$$
$$\begin{aligned}&\geq H([X_{(i,k_i)}]_{i\in I},[X_{(j_i,i)}]_{i\in I},Z_1,Z_2,[W_i]_{i\in I})\\&\quad+H([W_i]_{i\in I})\end{aligned} \tag{93}$$
$$\begin{aligned}&=H([X_{(i,k_i)}]_{i\in I},[X_{(j_i,i)}]_{i\in I},Z_1,Z_2,[W_i,W_{j_i},W_{k_i}]_{i\in I})\\&\quad+H([W_i]_{i\in I})\end{aligned} \tag{94}$$
$$\geq H([W_i,W_{j_i},W_{k_i}]_{i\in I})+H([W_i]_{i\in I})=2N_1F \tag{95}$$

where the (92) follows from that user 1 can decode $\{W_i:i\in I\}$ based on $\{X_{(i,k_i)}]_{i\in I},Z_1\}$ and user 2 can decode $\{W_i:i\in I\}$ based on $\{[X_{(j_i,i)}]_{i\in I},Z_2\}$; (93) follows from the matroidal Shannon inequality; (94) follows that user 1 can decode $\{W_{j_i}:i\in I\}$ based on $\{[X_{(j_i,i)}]_{i\in I},Z_1\}$ and user 2 can decode $\{W_{k_i}:i\in I\}$ based on $\{X_{(i,k_i)}]_{i\in I},Z_2\}$.

To balance the demand pairs $(d_1,d_2)\in[N_1]\times[N_2]$ uniformly, we introduce the the following cut-set bound for $l_i\in[N_2]$

$$M_1+\sum_{i\in[N_1]}R_{(i,l_i)}\geq N_1F. \tag{96}$$

The linear combination (95) + $\frac{(N_2-N_1)}{N_1}\times$ (96) yields

$$\begin{aligned}&\frac{N_2}{N_1}M_1+M_2+\sum_{i\in I}R_{(i,k_i)}+\sum_{i\in I}R_{(j_i,i)}\\&+\frac{N_2-N_1}{N_1}\sum_{i\in[N_1]}R_{(i,l_i)}\geq(N_1+N_2)F.\end{aligned} \tag{97}$$

Again we perform random assignment of the sets I, $\{j_i\}_{i\in I}$, $\{k_i\}_{i\in I}$, and $\{l_i\}_{i\in[N_1]}$ to construct the uniform average rate. We randomly choose $I=[N_1/2]$ or $I=[N_1]\backslash[N_1/2]$, each with probability $\frac{1}{2}$, $\{j_i\}=[N_1]\backslash I$, and randomly choose $\frac{N_1}{2}$ elements in $[N_2]\backslash[N_1]$ as $\{k_i\}$. Among the number of $\frac{N_2-N_1}{N_1}$ inequalities (96), we choose $\frac{1}{2}$ fraction of $\frac{(N_2-N_1)}{N_1}$ inequalities (96) to assign the rates $\{R_{(i,l_i)}\}$ such that $\{l_i\}\in I$ for $i\in I$ and for $\{l_i\}\in[N_1]\backslash I$ for $i\in[N_1]\backslash I$. On the other hand, we choose the remained $\frac{N_2-N_1}{N_1}-\frac{1}{2}$ fraction of $\frac{N_2-N_1}{N_1}$ inequalities (96) to assign the rates$\{R_{(i,l_i)}\}$ such that $l_i\in[N_2]\backslash[N_1]$ for $i\in[N_1]$. In this way, for any demand (d_1,d_2), $d_1,d_2\in[N_1]$, corresponding to the rate $R_{(d_1,d_2)}$ appeals with the probability

$$\frac{1}{2}\cdot\frac{N_1/2}{(N_1/2)^2}=\frac{1}{N_1}$$

and for any demand (d_1,d_2), $d_1\in[N_1]$, $d_2\in[N_2]\backslash[N_1]$, corresponding to the rate $R_{(d_1,d_2)}$ appeals with the probability

$$\frac{1}{2}\frac{\frac{N_1}{2}}{\frac{N_1}{2}(N_2-N_1)}+\left(\frac{N_2-N_1}{N_1}-\frac{1}{2}\right)\frac{N_1}{N_1(N_2-N_1)}=\frac{1}{N_1}.$$

That is, for all $(d_1,d_2)\in[N_1]\times[N_2]$, the rate $R_{(d_1,d_2)}$ appeals in (97) with probability $\frac{1}{N_1}$ such that we can replace the term $\sum_{i\in I}R_{(i,k_i)}+\sum_{i\in I}R_{(j_i,i)}+\frac{N_2-N_1}{N_1}\sum_{i\in[N_1]}R_{(i,l_i)}$ in (97) with $\frac{1}{N_1}\sum_{i\in[N_1],j\in[N_2]}R_{(i,j)}=N_2\tilde{R}$ or equivalently (P5+).

To prove (P6+), we consider N_1 even and $N_2\leq 2N_1$ such that $2N_2-2N_1\leq N_2$. Let the set $J\subseteq[N_2]$, $|J|=2(N_2-N_1)$, and $\{i_j\in[N_1]:j\in J\}$, we have the cut-set bound

$$M_2+\sum_{j\in J}R_{(i_j,j)}\geq 2(N_2-N_1)F. \tag{98}$$

The linear combination $2\times(95)+(98)$ yields

$$\begin{aligned}&2M_1+3M_2+2\sum_{i\in I}R_{(i,k_i)}+2\sum_{i\in I}R_{(j_i,i)}+\sum_{j\in J}R_{(i_j,j)}\\&\geq 2(N_1+N_2)F.\end{aligned} \tag{99}$$

Similarly we then perform random assignment of the sets I, $\{j_i\}_{i\in I}$, $\{k_i\}_{i\in I}$, J, and $\{i_j\}_{j\in J}$ to construct the uniform average rate. We choose $I=[N_1/2]$ and $I=[N_1]\backslash[N_1/2]$ for the two inequalities of (95), respectively, and randomly choose $N_1/2$ elements in $[N_2]\backslash[N_1]$ as $\{k_i\}_{i\in I}$ for each I. We construct set $J=[N_1]\cup A$, where A contains $2N_2-3N_1$

distinct elements randomly picked from $[N_2]\backslash[N_1]$. Then we assign $\{i_j : j = 1, \ldots, \frac{N_1}{2}\} = [N_1/2]$ with random permutation and assign $\{i_j : j = \frac{N_1}{2}+1, \ldots, N_1\} = [N_1]\backslash[N_1/2]$ with random permutation. Finally we randomly choose $2N_2 - 3N_1$ elements from $[N_1]$ to assign the rest of $\{i_j : j \in A\}$. In this way, for any demand (d_1, d_2), $d_1, d_2 \in [N_1]$, corresponding to the rate $R_{(d_1,d_2)}$ appeals with the probability

$$\frac{N_1/2}{(N_1/2)^2} = \frac{2}{N_1} \tag{100}$$

and for any demand (d_1, d_2), $d_1 \in [N_1]$, $d_2 \in [N_2]\backslash[N_1]$, corresponding to the rate $R_{(d_1,d_2)}$ appeals with the probability

$$\frac{N_1/2}{(N_1/2)(N_2 - N_1)} + \frac{2N_2 - 3N_1}{N_1(N_2 - N_1)} = \frac{2}{N_1}. \tag{101}$$

That is, for all $(d_1, d_2) \in [N_1] \times [N_2]$ the rate $R_{(d_1,d_2)}$ appeals in (99) with probability $\frac{2}{N_1}$ such that we can replace the term $2\sum_{i\in I} R_{(i,k_i)} + 2\sum_{i\in I} R_{(j_i,i)} + \sum_{j\in J} R_{(i_j,j)}$ in (99) with $\frac{2}{N_1}\sum_{i\in[N_1],j\in[N_2]} R_{(i,j)} = 2N_2\tilde{R}$ and hence we obtain (P6+).

To prove Proposition 7, we show that the vertices w_1 to w_7 in Fig. 5 are zero-error achievable. The vertices w_1 and w_2 can be achieved when in placement phase, user 1 caches $Z_1 = (W_1, \ldots, W_{N_1})$ of size N_1F and in delivery phase, for any demand $(d_1, d_2) \in [N_1] \times [N_2]$, the server transmits $X_{(d_1,d_2)} = W_{d_2}$ corresponding to the rate $R_{(d_1,d_2)} = F$ and hence the uniform-average rate is $\tilde{R} = F$. The vertices w_3 and w_2 can be achieved when in placement phase, user 1 caches $Z_1 = (W_1, \ldots, W_{N_1})$ of size N_1F and user 2 caches $Z_2 = (W_1, \ldots, W_{N_2})$ of size N_2F such that in delivery phase, no further transmission from server is required and hence the uniform-average rate is $\tilde{R} = 0$. The vertex w_5 can be achieved when in placement phase, user 2 caches $Z_2 = (W_1, \ldots, W_{N_2})$ of size N_2F and in delivery phase, for any demand $(d_1, d_2) \in [N_1] \times [N_2]$, the server transmits $X_{(d_1,d_2)} = W_{d_1}$ corresponding to the rate $R_{(d_1,d_2)} = F$ and hence the uniform-average rate is $\tilde{R} = F$.

The vertex w_6 can be achieved by first divide all the files W_i, $i \in [N_1]$, into two halves subfiles (U_i, V_i) of equal size $F/2$, i.e., $W_i = (U_i, V_i)$ for all $i \in [N_1]$. In the placement phase, user 1 caches $Z_1 = (U_1, \ldots, U_{N_1})$ of size $\frac{N_1}{2}F$ and user 2 caches $Z_2 = (V_1, \ldots, V_{N_1}, W_{N_1+1}, \ldots, W_{N_2})$ of size $\frac{N_1}{2}F + (N_2 - N_1)F = (N_1 - \frac{N_1}{2})F$. In the delivery phase for any demand (d_1, d_2), $d_1, d_2 \in [N_1]$, the server transmit $X_{(d_1,d_2)} = U_{d_2} \oplus V_{d_1}$ corresponding to rate $R_{(d_1,d_2)} = \frac{F}{2}$ and for any demand (d_1, d_2), $d_1 \in [N_1]$, $d_2 \in [N_2]\backslash[N_1]$, the server transmit $X_{(d_1,d_2)} = V_{d_1}$ corresponding to rate $R_{(d_1,d_2)} = \frac{F}{2}$ such that the uniform-average rate is $\tilde{R} = \frac{F}{2}$. The vertex w_7 can be achieved by first divide all the files W_i into two halves subfiles (U_i, V_i) of equal size $F/2$, i.e., $W_i = (U_i, V_i)$ for all $i \in [N_2]$. In the placement phase, user 1 caches $Z_1 = (U_1, \ldots, U_{N_2})$ of size $\frac{N_2}{2}F$ and user 2 caches $Z_2 = (V_1, \ldots, V_{N_2})$ of size $\frac{N_2}{2}F$. In the delivery phase for any demand (d_1, d_2), $d_1 \in [N_1]$, $d_2 \in [N_2]$, the server transmit $X_{(d_1,d_2)} = U_{d_2} \oplus V_{d_1}$ corresponding to rate $R_{(d_1,d_2)} = \frac{F}{2}$ and hence the uniform-average rate is $\tilde{R} = \frac{F}{2}$.

REFERENCES

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

[2] C. Tian, "Symmetry, demand types and outer bounds in caching systems," in *Proc. IEEE Int. Symp. Inform. Theory*, Jul. 2016, pp. 825–829.

[3] D. Cao, D. Zhang, P. Chen, N. Liu, W. Kang, and D. Gunduz, "Coded caching with heterogeneous cache sizes and link qualities: The two-user case," in *Proc. IEEE Int. Symp. Inform. Theory*, Jun. 2018, pp. 1545–1549.

[4] C. Li, "On rate region of caching problems with non-uniform file and cache sizes," *IEEE Commun. Lett.*, vol. 21, no. 2, pp. 238–241, Feb 2017.

[5] M. A. Maddah-Ali and U. Niesen, "Decentralized coded caching attains order-optimal memory-rate tradeoff," *IEEE/ACM Trans. Netw.*, vol. 23, no. 4, pp. 1029–1040, Aug 2015.

[6] J. Zhang, X. Lin, C. Wang, and X. Wang, "Coded caching for files with distinct file sizes," in *Proc. IEEE Int. Symp. Inform. Theory*, June 2015, pp. 1686–1690.

[7] Q. Yu, M. A. Maddah-Ali, and A. S. Avestimehr, "Characterizing the rate-memory tradeoff in cache networks within a factor of 2," *IEEE Trans. Inf. Theory*, vol. 65, no. 1, pp. 647–663, Jan 2019.

[8] U. Niesen and M. A. Maddah-Ali, "Coded caching with nonuniform demands," *IEEE Trans. Inf. Theory*, vol. 63, no. 2, pp. 1146–1158, Feb 2017.

[9] J. Zhang, X. Lin, and X. Wang, "Coded caching under arbitrary popularity distributions," *IEEE Trans. Inf. Theory*, vol. 64, no. 1, pp. 349–366, Jan. 2018.

[10] M. Ji, A. M. Tulino, J. Llorca, and G. Caire, "Order-optimal rate of caching and coded multicasting with random demands," *IEEE Trans. Inf. Theory*, vol. 63, no. 6, pp. 3923–3949, Jun. 2017.

[11] J. Hachem, N. Karamchandani, and S. N. Diggavi, "Coded caching for multi-level popularity and access," *IEEE Trans. Inf. Theory*, vol. 63, no. 5, pp. 3108–3141, May 2017.

[12] P. Quinton, S. Sahraei, and M. Gastpar, "A novel centralized strategy for coded caching with non-uniform demands," *arXiv:1801.10563*, 2018.

[13] A. M. Daniel and W. Yu, "Optimization of heterogeneous coded caching," *arXiv:1708.04322*, Aug. 2017.

[14] https://engineering.purdue.edu/~chihw/pub_pdf/ISIT2019_micro_cap_proof.pdf.