Device-to-Device Coded-Caching With Distinct Cache Sizes

Abdelrahman M. Ibrahim, *Member, IEEE*, Ahmed A. Zewail, *Member, IEEE*, and Aylin Yener, *Fellow, IEEE*

***Abstract*— This paper considers a cache-aided device-to-device (D2D) system where the users are equipped with cache memories of different size. During low traffic hours, a server places content in the users' cache memories, knowing that the files requested by the users during peak traffic hours will have to be delivered by D2D transmissions only. The worst-case D2D delivery load is minimized by jointly designing the uncoded cache placement and linear coded D2D delivery. Next, a novel lower bound on the D2D delivery load with uncoded placement is proposed and used in explicitly characterizing the minimum D2D delivery load (MD2DDL) with uncoded placement for several cases of interest. In particular, having characterized the MD2DDL for equal cache sizes, it is shown that the same delivery load can be achieved in the network with users of unequal cache sizes, provided that the smallest cache size is greater than a certain threshold. The MD2DDL is also characterized in the small cache size regime, the large cache size regime, and the three-user case. Comparisons of the server-based delivery load with the D2D delivery load are provided. Finally, connections and mathematical parallels between cache-aided D2D systems and coded distributed computing (CDC) systems are discussed.**

***Index Terms*— Coded caching, uncoded placement, device-to-device communication, unequal cache sizes.**

I. Introduction

DEVELOPMENT of novel techniques that fully utilize network resources is imperative to meet the objectives of 5G systems and beyond with increasing demand for wireless data traffic, e.g., video-on-demand services [1]. Device-to-device (D2D) communications [2] and caching [3] are two prominent techniques for alleviating network congestion. D2D communications utilize the radio interface enabling the nodes to directly communicate with each other to reduce the delivery load on servers/base stations/access points. Caching schemes utilize the nodes' cache memories to shift some of the network traffic to low congestion periods. In coded caching [4], the server jointly designs the content placement during off-peak hours and the content delivery during peak hours, to create multicast coding opportunities. That is, coded caching not only shifts the network traffic to off-peak hours but also creates multicast opportunities that reduce the delivery load on the server [4]. In particular, in the placement phase, the server first partitions the files into pieces. Then, the server either places uncoded or coded pieces of the files at the users' cache memories. Most of the work on coded caching considers uncoded placement [4]–[15], for its practicality and near optimality [7]–[9]. References [8], [9] have illustrated that the server-based delivery problem in [4] is equivalent to an index-coding problem and the delivery load in [4] is lower bounded by the acyclic index-coding bound [16, Corollary 1]. Reference [7] has proposed an alternative proof for the uncoded placement bound [8], [9] using a genie-aided approach.

Coded caching in device-to-device networks has been investigated in [6], [17]–[24]. In particular, D2D coded caching was first considered in [6], where centralized and decentralized caching schemes have been proposed for when the users have equal cache sizes. References [6], [17]–[19] have studied the impact of coded caching on throughput scaling laws of D2D networks under the protocol model in [25]. Reference [20] has considered a D2D system where only a subset of the users participate in delivering the missing subfiles to all users. Reference [21] has proposed using random linear network coding to reduce the delay experienced by the users in lossy networks. Reference [22] has proposed a secure D2D delivery scheme that protects the D2D transmissions in the presence of an eavesdropper. Reference [23] has considered secure D2D coded caching when each user can recover its requested file and is simultaneously prevented from accessing any other file.

More realistic caching models that reflect the heterogeneity in content delivery networks consider systems with distinct cache sizes [10]–[15], [26]–[28], unequal file sizes [27], [29], [30], distinct distortion requirements [26], [31], [32], and non-uniform popularity distributions [33]–[38]. In this work, we focus on the distinct cache sizes, i.e., the varying storage capabilities of the users. This setup has been considered in [10]–[15], [26]–[28] for the server-based delivery problem of [4]. In particular, in [13], [15], we have shown that the delivery load is minimized by solving a linear program over the parameters of the uncoded placement and linear delivery schemes.

Manuscript received March 8, 2019; revised August 30, 2019 and November 29, 2019; accepted January 21, 2020. Date of publication January 31, 2020; date of current version May 15, 2020. This work was supported in part by NSF under Grant CCF-1749665. This research was performed when the authors were with the department of Electrical Engineering, Pennsylvania State University, University Park, PA 16802 USA. This article was presented in part at the 2018 IEEE International Conference on Communications (ICC), Kansas City, MO, USA. The associate editor coordinating the review of this article and approving it for publication was S. Mohajer. *(Corresponding author: Aylin Yener.)*

Abdelrahman M. Ibrahim and Ahmed A. Zewail were with the Department of Electrical Engineering, Pennsylvania State University, University Park, PA 16802 USA. They are now with the Department of Wireless Research and Development, Qualcomm Technologies Inc., San Diego, CA 92121 USA (e-mail: abdelrah@qti.qualcomm.com; azewail@qti.qualcomm.com).

Aylin Yener was with the Department of Electrical Engineering, Pennsylvania State University, University Park, PA 16802 USA. She is now with the Department of Electrical and Computer Engineering, The Ohio State University, Columbus, OH 43210 USA (e-mail: yener@ee.psu.edu).

Color versions of one or more of the figures in this article are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TCOMM.2020.2970950

Different from [13], [15] and all references with distinct cache sizes, in this paper, we investigate coded caching with end-users of unequal cache sizes when the delivery phase must be carried out by D2D transmissions. That is, the placement and delivery design must be such that the server does not participate in delivery at all, thus saving its resources to serve those outside the D2D network. This distinction calls for new placement and delivery schemes as compared to serve-based delivery architectures [10]–[15], [26]–[28]. In the same spirit as [15], we show that a linear program minimizes the D2D delivery load by optimizing over the partitioning of the files in the placement phase and the size and structure of the D2D transmissions, and find the optimal design. We remark that even though the proposed optimization framework is inspired by our work in [15], finding device-to-device delivery schemes with optimization constraints is a non-trivial extension of [15] due to the inherent design flexibility and unique delivery challenges in D2D systems. In the D2D setting, the transmissions from all users are jointly optimized in order to minimize the total D2D delivery load. In addition, we show that the trade-off between the delivery load and the cache sizes has characteristics that are unique to the D2D setting that could not have been addressed by the centralized formulation. For example, in the D2D setting the heterogeneity in users cache sizes does not lead to an increase in the achievable D2D delivery load as long as the smallest cache is large enough. This work also explains the relationship between the server-based and D2D delivery problems, which has not been addressed in previous works.

Building on the techniques in [7]–[9], we derive a lower bound on the worst-case D2D delivery load with uncoded placement and one-shot delivery [24], which is also defined by a linear program. Using the proposed lower bound, we first prove the optimality of the caching scheme in [6] assuming uncoded placement and one-shot delivery for systems with equal cache sizes. Next, we explicitly characterize the D2D delivery load memory trade-off assuming uncoded placement and one-shot delivery for several cases of interest. In particular, we show that the D2D delivery load depends only on the total cache size in the network whenever the smallest cache size is greater than a certain threshold. For a small system with three users, we identify the precise trade-off for any library size. For larger systems, we characterize the trade-off in two regimes, i.e., the small total cache size regime and in the large total cache size regime, which are defined in the sequel. For remaining sizes of the total network cache, we observe numerically that the proposed caching scheme achieves the minimum D2D delivery load assuming uncoded placement. Finally, we establish the relationship between the server-based and D2D delivery loads assuming uncoded placement. We also discuss the parallels between the recent coded distributed computing (CDC) framework [39] and demonstrate how it relates to D2D caching systems.

The remainder of this paper is organized as follows. In Section II, we describe the system model and the main assumptions. The optimization problems characterizing the upper and lower bounds on the minimum D2D delivery load are formulated in Section III-A. Section III-B summarizes our results on the minimum D2D delivery load with uncoded placement. The general caching scheme is developed in Section IV. Section V explains the caching schemes that achieve the D2D delivery loads presented in Section III-B. The optimality of uncoded placement is investigated in Section VI. In Section VII, we discuss the trade-off in the general case, the connection to server-based systems, and connections to distributed computing. Section VIII provides the conclusions.

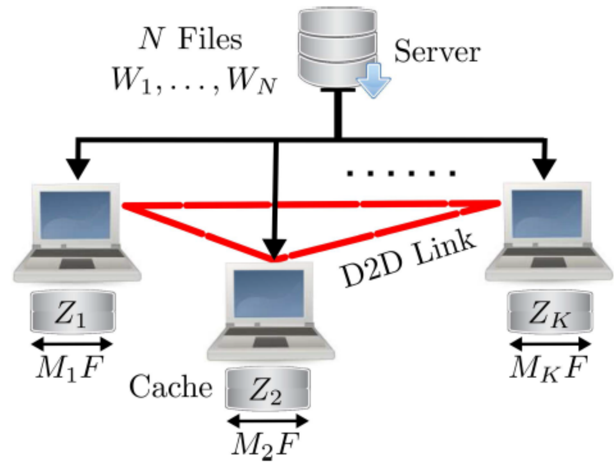


Fig. 1. D2D caching with unequal cache sizes at the end-users.

II. System Model

Notation: Vectors are represented by boldface letters, $\oplus$ refers to bitwise XOR operation, $|W|$ denotes size of W, $\mathcal{A} \setminus \mathcal{B}$ denotes the set of elements in $\mathcal{A}$ and not in $\mathcal{B}$, $[K] \triangleq \{1, \ldots, K\}$, ϕ denotes the empty set, $\subsetneq_\phi [K]$ denotes non-empty subsets of $[K]$, and $\mathcal{P}_\mathcal{A}$ is the set of all permutations of the elements in the set $\mathcal{A}$, e.g., $\mathcal{P}_{\{1,2\}} = \{[1, 2],\ [2, 1]\}$.

Consider a server connected to K users via a shared error-free link, and the users are connected to each other via error-free device-to-device (D2D) communication links, as illustrated in Fig. 1. The server has a library of N files, $W_1, \ldots, W_N$, each with size F bits. End-users are equipped with cache memories that have different sizes, the size of the cache memory at user k is equal to M_kF bits. Without loss of generality, let $M_1 \leq M_2 \leq \cdots \leq M_K$. Define m_k to denote the memory size of user k normalized by the library size NF, i.e., $m_k = M_k/N$. Let $\boldsymbol{M} = [M_1, \ldots, M_K]$ and $\boldsymbol{m} = [m_1, \ldots, m_K]$. We focus on the more practical case where the number of users is less than the number of files, i.e., $K \leq N$, e.g., a movie database serving cooperative users in a 5G hybrid cloud-fog access network [40].

D2D caching systems operate similarly to server-based systems in the placement phase, but differ in the delivery phase. Namely, in the placement phase, the server designs the users' cache contents without knowing their demands and knowing that it will not participate in the delivery phase. The content of the cache at user k is denoted by Z_k and satisfies the size constraint $|Z_k| \leq M_kF$ bits. Formally, Z_k is defined as follows.

Definition 1 (Cache Placement): A cache placement function $\phi_k : [2^F]^N \to [2^F]^{M_k}$ maps the files in the library to the cache memory of user k, i.e., $Z_k = \phi_k(W_1, W_2, .., W_N)$.

Just before the delivery phase, users announce their file demands. The demand vector is denoted by $\boldsymbol{d} = [d_1, \dots, d_K]$ such that W_{d_k} is the file requested by user k. The requested files must be delivered by utilizing D2D communications only [6], which requires that the sum of the users' cache sizes is at least equal to the library size, i.e., $\sum_{k=1}^{K} m_k \geq 1$. More specifically, user j transmits the sequence of unicast/multicast signals, $X_{j\to\mathcal{T},\boldsymbol{d}}$, to the users in the set $\mathcal{T} \subsetneq_\phi [K]\setminus\{j\}$. Let $|X_{j\to\mathcal{T},\boldsymbol{d}}| = v_{j\to\mathcal{T}}F$ bits, i.e., the *transmission variable* $v_{j\to\mathcal{T}} \in [0,1]$ represents the amount of data delivered to the users in $\mathcal{T}$ by user j as a fraction of the file size F.

Definition 2 (Encoding): Given demand $\boldsymbol{d}$*, an encoding* $\psi_{j\to\mathcal{T}} : [2^F]^{M_j} \times [N]^K \to [2^F]^{v_{j\to\mathcal{T}}}$ *maps the content cached by user* j *to a signal sent to the users in* $\mathcal{T} \subsetneq_\phi [K]\setminus\{j\}$*, i.e., the signal* $X_{j\to\mathcal{T},\boldsymbol{d}} = \psi_{j\to\mathcal{T}}(Z_j, \boldsymbol{d})$ *and* $|X_{j\to\mathcal{T},\boldsymbol{d}}| = v_{j\to\mathcal{T}}F$.

At the end of the delivery phase, user k must be able to reconstruct W_{d_k} reliably using the received D2D signals $\{X_{j\to\mathcal{T},\boldsymbol{d}}\}_{j\neq k,\mathcal{T}}$ and its cache content Z_k. Let $R_j \triangleq \sum_{\mathcal{T}\subsetneq_\phi [K]\setminus\{j\}} v_{j\to\mathcal{T}}$ be the amount of data transmitted by user j, normalized by the file size F.

Definition 3 (Decoding): Given the demand $\boldsymbol{d}$*, a decoding function* $\mu_k : [2^F]^{\sum_{j\neq k} R_j} \times [2^F]^{M_k} \times [N]^K \to [2^F]$*, maps the D2D signals* $X_{j\to\mathcal{T},\boldsymbol{d}}, \forall j \in [K]\setminus\{k\}, \mathcal{T} \subsetneq_\phi [K]\setminus\{j\}$ *and the content cached by user* k *to* $\hat{W}_{d_k}$*, i.e.,* $\hat{W}_{d_k} = \mu_k\left(\{X_{j\to\mathcal{T},\boldsymbol{d}}\}_{j\neq k,\mathcal{T}}, Z_k, \boldsymbol{d}\right)$.

The achievable D2D delivery load is defined as follows.

Definition 4: For a given $\boldsymbol{m}$*, the D2D delivery load* $R(\boldsymbol{m}) \triangleq \sum_{j=1}^{K} R_j(\boldsymbol{m})$ *is said to be achievable if for every* $\epsilon > 0$ *and large enough* F*, there exists* $(\phi_k(.), \psi_{j\to\mathcal{T}}(.), \mu_k(.))$ *such that* $\max_{\boldsymbol{d},k\in[K]} Pr(\hat{W}_{d_k} \neq W_{d_k}) \leq \epsilon$*, and* $R^*(\boldsymbol{m}) \triangleq \inf\{R : R(\boldsymbol{m})$ *is achievable*$\}$.

In general, an achievable D2D delivery scheme satisfies the decodability constraints

$$H\left(W_{d_k} \middle| \{X_{j\to\mathcal{T},\boldsymbol{d}}\}_{j\neq k,\mathcal{T}}, Z_k\right) = 0, \quad \forall k. \tag{1}$$

In this work, we focus on one-shot delivery schemes [24] where W_{d_k} is partitioned into $W_{d_k}^{(1)}, \dots, W_{d_k}^{(K)}$, such that $W_{d_k}^{(k)}$ is cached by user k and $W_{d_k}^{(j)}$ is decoded using the transmissions from user j only. That is, we have the following decodability constraints

$$H\left(W_{d_k}^{(j)} \middle| \{X_{j\to\mathcal{T},\boldsymbol{d}}\}_{\mathcal{T}}, Z_k\right) = 0, \quad \forall j \neq k, \forall k. \tag{2}$$

Similar to much of the coded caching literature [4]–[15], [27], we will consider placement schemes where the users cache only pieces of the files, i.e., uncoded placement. We denote the set of such schemes with $\mathfrak{A}$. In the delivery phase, we consider the class of delivery policies $\mathfrak{D}$, which is based on interference cancellation. In particular, we consider clique-covering schemes [8] where users generate the multicast signals with XORed pieces of files such that each user $k \in \mathcal{T}$ cancels the interference from $X_{j\to\mathcal{T},\boldsymbol{d}}$ in order to decode its desired piece. For a caching scheme in $(\mathfrak{A}, \mathfrak{D})$, we define the following.

Definition 5: For an uncoded placement scheme in $\mathfrak{A}$*, and a delivery policy in* $\mathfrak{D}$*, the achievable worst-case D2D delivery load is defined as*

$$R_{\mathfrak{A},\mathfrak{D}} \triangleq \max_{\boldsymbol{d}\in[N]^K} \sum_{j=1}^{K} R_{j,\boldsymbol{d},\mathfrak{A},\mathfrak{D}} = \sum_{j=1}^{K} \sum_{\mathcal{T}\subsetneq_\phi [K]\setminus\{j\}} v_{j\to\mathcal{T}}, \tag{3}$$

and $R^*_{\mathfrak{A},\mathfrak{D}}$ *denotes the minimum delivery load achievable with a caching scheme in* $(\mathfrak{A}, \mathfrak{D})$.

Definition 6: For an uncoded placement scheme in $\mathfrak{A}$ *and any one-shot delivery scheme,*

$$R^*_{\mathfrak{A}}(\boldsymbol{m}) \triangleq \inf\{R_{\mathfrak{A}} : R_{\mathfrak{A}}(\boldsymbol{m}) \text{ is achievable}\}, \tag{4}$$

is the minimum D2D delivery load achievable with uncoded placement and one-shot delivery.

III. Main Results

In this work, we propose a caching scheme where the cache placement is paramterized by the allocation vector $\boldsymbol{a}$, such that the allocation variable $a_{\mathcal{S}}$ determines the size of the subfile stored exclusively at the users in $\mathcal{S}$. The proposed delivery procedure is parameterized by the vectors $\boldsymbol{v}$ and $\boldsymbol{u}$, where the former determines the size of the transmitted signals $X_{j\to\mathcal{T}}$ and the latter specifies the structure of the transmitted signals. In Theorem 1, we optimize over the parameters of the proposed caching scheme in order to minimize the D2D delivery load.

Next, we illustrate the proposed caching scheme with an example. We consider a case where the heterogeneity in cache sizes does not increase the delivery load, i.e., we achieve the same delivery load in a homogeneous system with the same aggregate cache size. More sepcifically, the delivery scheme in [6] can be generalized for unequal cache sizes, by considering D2D transmissions with different sizes.

Example 1: For $K = N = 3$ *and* $\boldsymbol{m} = [0.6, 0.7, 0.8]$*, the proposed caching scheme is as follows:*

Placement Phase: *Each file* W_n *is divided into subfiles* $\tilde{W}_{n,\{1,2\}}, \tilde{W}_{n,\{1,3\}}, \tilde{W}_{n,\{2,3\}}, \tilde{W}_{n,\{1,2,3\}}$*, where* $\tilde{W}_{n,\mathcal{S}}$ *is stored exclusively at the users in* $\mathcal{S}$*, e.g.,* $\tilde{W}_{n,\{1,2\}}$ *is stored at users* $\{1,2\}$*. We assume* $|\tilde{W}_{n,\mathcal{S}}| = a_{\mathcal{S}}F, \forall n$*. More specifically,* $a_{\{1,2\}} = 0.2$*,* $a_{\{1,3\}} = 0.3$*,* $a_{\{2,3\}} = 0.4$*, and* $a_{\{1,2,3\}} = 0.1$.

Delivery Phase: *User 1 sends* $X_{1\to\{2,3\}} = W^{1\to\{2,3\}}_{d_2,\{1,3\}} \oplus W^{1\to\{2,3\}}_{d_3,\{1,2\}}$*, where* $W^{1\to\{2,3\}}_{d_2,\{1,3\}} \subset \tilde{W}_{d_2,\{1,3\}}$*,* $W^{1\to\{2,3\}}_{d_3,\{1,2\}} \subset \tilde{W}_{d_3,\{1,2\}}$*. Similarly, we have* $X_{2\to\{1,3\}} = W^{2\to\{1,3\}}_{d_1,\{2,3\}} \oplus W^{2\to\{1,3\}}_{d_3,\{1,2\}}$ *and* $X_{3\to\{1,2\}} = W^{3\to\{1,2\}}_{d_1,\{2,3\}} \oplus W^{3\to\{1,2\}}_{d_2,\{1,3\}}$*. We assume* $|W^{j\to\mathcal{T}}_{d_k,\mathcal{S}}| = u^{j\to\mathcal{T}}_{\mathcal{S}}F$*. More specifically, we have*

- $|X_{1\to\{2,3\}}|/F = v_{1\to\{2,3\}} = u^{1\to\{2,3\}}_{\{1,2\}} = u^{1\to\{2,3\}}_{\{1,3\}} = 0.05$.
- $|X_{2\to\{1,3\}}|/F = v_{2\to\{1,3\}} = u^{2\to\{1,3\}}_{\{1,2\}} = u^{2\to\{1,3\}}_{\{2,3\}} = 0.15$.
- $|X_{3\to\{1,2\}}|/F = v_{3\to\{1,2\}} = u^{3\to\{1,2\}}_{\{1,3\}} = u^{3\to\{1,2\}}_{\{2,3\}} = 0.25$.

The placement and delivery phases are illustrated in Fig. 2. Note that the same delivery load is achieved by the caching scheme in [6] for $\boldsymbol{m} = [0.7, 0.7, 0.7]$*. In Theorem 7, we show that the proposed scheme achieves* $R^*_{\mathfrak{A}}(\boldsymbol{m}) = 3/2 - (m_1 + m_2 + m_3)/2 = 0.45$.

A. Performance Bounds

First, we have the following parameterization for the optimum of the class of caching schemes under consideration.

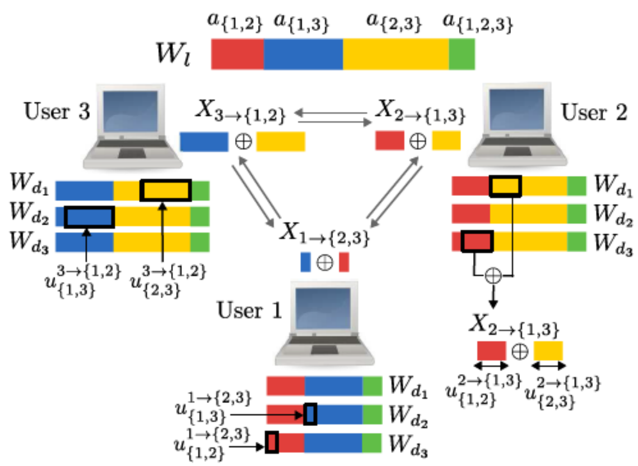


Fig. 2. Example $K = N = 3$, and $\boldsymbol{m} = [0.6, 0.7, 0.8]$.

Theorem 1: Given $N \geq K$, and $\boldsymbol{m}$, the minimum worst-case D2D delivery load assuming uncoded placement and a delivery policy in $\mathfrak{D}$, $R^*_{\mathfrak{A},\mathfrak{D}}(\boldsymbol{m})$, is characterized by the following linear program

$$\textbf{\textit{O1:}}\ R^*_{\mathfrak{A},\mathfrak{D}}(\boldsymbol{m}) = \min_{\boldsymbol{a},\boldsymbol{u},\boldsymbol{v}} \sum_{j=1}^{K} \sum_{\mathcal{T}\subsetneq_\phi [K]\setminus\{j\}} v_{j\to\mathcal{T}} \tag{5a}$$

$$\text{subject to } \boldsymbol{a} \in \mathfrak{A}(\boldsymbol{m}), \tag{5b}$$

$$(\boldsymbol{u}, \boldsymbol{v}) \in \mathfrak{D}(\boldsymbol{a}), \tag{5c}$$

where $\mathfrak{A}(\boldsymbol{m})$ is set of uncoded placement schemes defined as

$$\mathfrak{A}(\boldsymbol{m}) = \Bigg\{ \boldsymbol{a} \in [0,1]^{2^K} \Bigg| \sum_{\mathcal{S}\subsetneq_\phi [K]} a_{\mathcal{S}} = 1, \sum_{\mathcal{S}\subset[K]\,:\,k\in\mathcal{S}} a_{\mathcal{S}} \leq m_k, \forall k \in [K] \Bigg\}, \tag{6}$$

and $\mathfrak{D}(\boldsymbol{a})$ is the set of feasible delivery schemes defined by

$$v_{j\to\{i\}} = a_{\{j\}} + \sum_{\mathcal{S}\subset[K]\setminus\{i\}\,:\,j\in\mathcal{S},|\mathcal{S}|\geq 2} u^{j\to\{i\}}_{\mathcal{S}}, \ \forall j \in [K],\ \forall i \in \mathcal{T}, \tag{7}$$

$$v_{j\to\mathcal{T}} = \sum_{\mathcal{S}\in\mathcal{B}_i^{j\to\mathcal{T}}} u^{j\to\mathcal{T}}_{\mathcal{S}}, \quad \forall j \in [K],\ \forall \mathcal{T} \subsetneq_\phi [K]\setminus\{j\},\ \forall i \in \mathcal{T}, \tag{8}$$

$$\sum_{j\in\mathcal{S}} \sum_{\mathcal{T}\subset\{i\}\cup(\mathcal{S}\setminus\{j\})\,:\,i\in\mathcal{T}} u^{j\to\mathcal{T}}_{\mathcal{S}} = a_{\mathcal{S}}, \quad \forall i \notin \mathcal{S}, \quad \forall \mathcal{S} \subset [K] \text{ s.t. } 2 \leq |\mathcal{S}| \leq K-1, \tag{9}$$

$$0 \leq u^{j\to\mathcal{T}}_{\mathcal{S}} \leq a_{\mathcal{S}}, \quad \forall j \in [K],\ \forall \mathcal{T} \subsetneq_\phi [K]\setminus\{j\},\ \forall \mathcal{S} \in \bigcup_{i\in\mathcal{T}} \mathcal{B}_i^{j\to\mathcal{T}}, \tag{10}$$

where $\mathcal{B}_i^{j\to\mathcal{T}} \triangleq \Big\{ \mathcal{S} \subset [K]\setminus\{i\} : \{j\}\cup(\mathcal{T}\setminus\{i\}) \subset \mathcal{S} \Big\}$.

Proof: Proof is provided in Section IV. □

Motivated by the lower bounds on server-based delivery in [7]–[9], we next establish that the minimum D2D delivery load memory trade-off with uncoded placement, $R^*_{\mathfrak{A}}(\boldsymbol{m})$, is lower bounded by the linear program defined in Theorem 2.

Theorem 2: Given $N \geq K$, and $\boldsymbol{m}$, the minimum worst-case D2D delivery load with uncoded placement and one-shot delivery, $R^*_{\mathfrak{A}}(\boldsymbol{m})$, is lower bounded by

$$\textbf{\textit{O2:}} \max_{\lambda_0\in\mathbb{R},\lambda_k\geq 0,\alpha_q\geq 0} \ -\lambda_0 - \sum_{k=1}^{K} m_k \lambda_k \tag{11a}$$

$$\text{subject to } \lambda_0 + \sum_{k\in\mathcal{S}} \lambda_k + \gamma_{\mathcal{S}} \geq 0, \quad \forall \mathcal{S} \subsetneq_\phi [K], \tag{11b}$$

$$\sum_{\boldsymbol{q}\in\mathcal{P}_{[K]\setminus\{j\}}} \alpha_{\boldsymbol{q}} = 1, \quad \forall j \in [K], \tag{11c}$$

where $\mathcal{P}_{[K]\setminus\{j\}}$ is the set of all permutations of the users in $[K]\setminus\{j\}$, $\alpha_{\boldsymbol{q}}$ are the coefficients of the convex combination over all $\boldsymbol{q} \in \mathcal{P}_{[K]\setminus\{j\}}$, and

$$\gamma_{\mathcal{S}} \triangleq \begin{cases} K-1, \text{ for } |\mathcal{S}| = 1, \\ \min_{j\in\mathcal{S}} \Bigg\{ \sum_{i=1}^{K-|\mathcal{S}|} \sum_{\substack{\boldsymbol{q}\in\mathcal{P}_{[K]\setminus\{j\}}:\ q_{i+1}\in\mathcal{S}, \\ \{q_1,\dots,q_i\}\cap\mathcal{S}=\phi}} i\,\alpha_{\boldsymbol{q}} \Bigg\}, \text{ for } 2 \leq |\mathcal{S}| \leq K-1 \\ 0, \text{ for } \mathcal{S} = [K]. \end{cases} \tag{12}$$

Proof: The proof is detailed in Section VI-A. □

B. Explicit Characterization Results

Next, using Theorems 1 and 2, we characterize the trade-off explicitly for several cases, which are illustrated in Table I. In particular, for these cases we show that $R^*_{\mathfrak{A}}(\boldsymbol{m}) = R^*_{\mathfrak{A},\mathfrak{D}}(\boldsymbol{m})$. First, using Theorem 2, we show the optimality of the D2D caching scheme proposed in [6] for systems where the users have equal cache sizes.

Theorem 3: For $N \geq K$, and $m_k = m = t/K, t \in [K]$, $\forall k \in [K]$, the minimum worst-case D2D delivery load with uncoded placement and one-shot delivery, $R^*_{\mathfrak{A}}(\boldsymbol{m}) = (1-m)/m$. In general, we have

$$R^*_{\mathfrak{A}}(\boldsymbol{m}) = \left(\frac{K-t}{t}\right)(t+1-Km) + \left(\frac{K-t-1}{t+1}\right)(Km-t), \tag{13}$$

where $t \in [K-1]$ and $t \leq Km \leq t+1$.

Proof: **Achievability**: The D2D caching scheme proposed in [6] achieves (13), which is also the optimal solution of (5). **Converse**: The proof is detailed in Section VI-B. □

Next theorem shows that the heterogeneity in users cache sizes does not increase the achievable D2D delivery load as long as the smallest cache m_1 is large enough.

Theorem 4: For $N \geq K$, $m_1 \leq \dots \leq m_K$, and $m_1 \geq \big(\sum_{k=2}^{K} m_k - 1\big)/(K-2)$, the minimum worst-case D2D delivery load with uncoded placement and one-shot delivery,

$$R^*_{\mathfrak{A}}(\boldsymbol{m}) = \left(\frac{K-t}{t}\right)\Big(t+1-\sum_{k=1}^{K} m_k\Big) + \left(\frac{K-t-1}{t+1}\right)\Big(\sum_{k=1}^{K} m_k - t\Big), \tag{14}$$

where $t \leq \sum_{k=1}^{K} m_k \leq t+1$, and $t \in [K-1]$.

TABLE I

SUMMARY OF THE ANALYTICAL RESULTS ON $R^*_{\mathfrak{A}}(\boldsymbol{m})$

Regions	$\sum_{j=1}^{K} m_j \in [1,2]$	$\dots$	$\sum_{j=1}^{K} m_j \in [t,t+1]$	$\dots$	$\sum_{j=1}^{K} m_j \in [K-1,K]$
$(K-2)m_1 \geq \sum_{j=2}^{K} m_j - 1$	Exact (14)	$\dots$	Exact (14)	$\dots$	Exact (14)
$(K-2)m_1 < \sum_{j=2}^{K} m_j - 1$ $(K-3)m_2 \geq \sum_{j=3}^{K} m_j - 1$	Exact (15)		Achievability (5) Lower bound (11)		Exact (16)
$\vdots$		$\dots$		$\dots$	
$(K-l-1)m_l < \sum_{j=l+1}^{K} m_j - 1$ $(K-l-2)m_{l+1} \geq \sum_{j=l+2}^{K} m_j - 1$					
$\vdots$					
$m_{K-2} < m_{K-1} + m_K - 1$					

Proof: **Achievability**: In Section V-A, we generalize the caching scheme in [6] to accommodate the heterogeneity in cache sizes. **Converse**: The proof is detailed in Section VI-C. □

The next theorem characterizes the trade-off in the small memory regime defined as the total network cache memory is less than twice the library size.

Theorem 5: For $N \geq K$, $m_1 \leq \dots \leq m_K$, $1 \leq \sum_{k=1}^{K} m_k \leq 2$, the minimum worst-case D2D delivery load with uncoded placement and one-shot delivery,

$$R^*_{\mathfrak{A}}(\boldsymbol{m}) = \frac{3K-l-2}{2} - \sum_{i=1}^{l} (K-i) m_i - \left(\frac{K-l}{2}\right) \sum_{i=l+1}^{K} m_i, \quad (15)$$

where l is an integer in $[K-2]$ such that $m_l < \frac{\sum_{i=l+1}^{K} m_i - 1}{K-l-1}$ and $m_{l+1} \geq \frac{\sum_{i=l+2}^{K} m_i - 1}{K-l-2}$.

Proof: **Achievability**: The caching scheme is provided in Section V-B. **Converse**: The proof is detailed in Section VI-D. □

From (15), we observe that the trade-off in the lth heterogeneity level depends on the individual cache sizes of users $\{1,\dots,l\}$ and the total cache sizes of the remaining users.

Remark 1: The trade-off in the region where $\sum_{k=1}^{K} m_k \leq 2$ and $(K-2)\ m_1 \geq \sum_{i=2}^{K} m_i - 1$, which is included in Theorem 4, can also be obtained by substituting $l = 0$ in Theorem 5.

The next theorem characterizes the trade-off in the large memory regime defined as one where the total network memory satisfies $\sum_{k=1}^{K} m_k \geq K-1$. In particular, we show the optimality of uncoded placement and one-shot delivery, i.e., $R^*_{\mathfrak{A}}(\boldsymbol{m}) = R^*(\boldsymbol{m})$.

Theorem 6: For $N \geq K$, $m_1 \leq \dots \leq m_K$, and $\sum_{k=1}^{K} m_k \geq K-1$, the minimum worst-case D2D delivery load with uncoded placement and one-shot delivery,

$$R^*_{\mathfrak{A}}(\boldsymbol{m}) = R^*(\boldsymbol{m}) = 1 - m_1, \quad (16)$$

where $m_1 < \frac{\sum_{i=2}^{K} m_i - 1}{K-2}$.

Proof: **Achievability**: The caching scheme is provided in Section V-C. **Converse**: The proof follows from the cut-set bound in [6]. □

Finally, for $K = 3$, we have the complete characterization below.

Theorem 7: For $K = 3$, $N \geq 3$, and $m_1 \leq m_2 \leq m_3$, the minimum worst-case D2D delivery load with uncoded placement and one-shot delivery,

$$R^*_{\mathfrak{A}}(\boldsymbol{m}) = \max\left\{\frac{7}{2} - \frac{3}{2}(m_1+m_2+m_3), 3-2m_1-m_2-m_3, \frac{3}{2} - \frac{1}{2}(m_1+m_2+m_3), 1-m_1\right\}. \quad (17)$$

Proof: **Achievability**: The proof is in Appendix A. **Converse**: The proof is in Appendix B. □

IV. GENERAL CACHING SCHEME

In the placement phase, we consider all feasible uncoded placement schemes in which the whole library can be retrieved utilizing the users' cache memories via D2D delivery, i.e., there must be no subfile stored at the server that is not placed in the end nodes in pieces. The delivery phase consists of K transmission stages, in each of which one of the K users acts as a "server". In particular, in the jth transmission stage, user j transmits the signals $X_{j \to \mathcal{T}}$ to the users in the sets $\mathcal{T} \subsetneq_{\phi} [K] \setminus \{j\}$.[1]

A. Placement Phase

The server partitions each file W_n into $2^K - 1$ subfiles, $\tilde{W}_{n,\mathcal{S}}, \mathcal{S} \subsetneq_{\phi} [K]$, such that $\tilde{W}_{n,\mathcal{S}}$ denotes a subset of

[1]For convenience, we omit the subscript $\boldsymbol{d}$ from $X_{j \to \mathcal{T}, \boldsymbol{d}}$ whenever the context is clear.

W_n which is stored exclusively at the users in the set $\mathcal{S}$. The partitioning is symmetric over the files, i.e., $|\tilde{W}_{n,\mathcal{S}}| = a_{\mathcal{S}}F$ bits, $\forall n \in [N]$, where the *allocation variable* $a_{\mathcal{S}} \in [0,1]$ defines the size of $\tilde{W}_{n,\mathcal{S}}$ as a fraction of the file size F. Therefore, the set of feasible uncoded placement schemes, $\mathfrak{A}(m)$, is defined by

$$\mathfrak{A}(m) = \Bigg\{ a \in [0,1]^{2^K} \Bigg| \sum_{\mathcal{S} \subsetneq_{\phi} [K]} a_{\mathcal{S}} = 1, \sum_{\mathcal{S} \subset [K] \,:\, k \in \mathcal{S}} a_{\mathcal{S}} \le m_k, \forall k \in [K] \Bigg\}, \tag{18}$$

where the allocation vector a consists of the allocation variables $a_{\mathcal{S}}, \mathcal{S} \subsetneq_{\phi} [K]$, the first constraint follows from the fact the whole library can be reconstructed from the users' cache memories, and the second represents the cache size constraint at user k. More specifically, user k cache content is defined as

$$Z_k = \bigcup_{n \in [N]} \bigcup_{\mathcal{S} \subset [K] \,:\, k \in \mathcal{S}} \tilde{W}_{n,\mathcal{S}}. \tag{19}$$

Next, we explain the delivery scheme for a three-user system for clarity of exposition, then we generalize to $K > 3$.

B. Delivery Phase: Three-User System

1) Structure of $X_{j \to \mathcal{T}}$: In the first transmission stage, i.e., $j = 1$, user 1 transmits the unicast signals $X_{1 \to \{2\}}, X_{1 \to \{3\}}$, and the multicast signal $X_{1 \to \{2,3\}}$ to users $\{2,3\}$. In particular, the unicast signal $X_{1 \to \{2\}}$ delivers the subset of W_{d_2} which is stored exclusively at user 1, i.e., subfile $\tilde{W}_{d_2,\{1\}}$, in addition to a fraction of the subfile stored exclusively at users $\{1,3\}$, which we denote by $W^{1 \to \{2\}}_{d_2,\{1,3\}}$. In turn, $X_{1 \to \{2\}}$ is given by

$$X_{1 \to \{2\}} = \tilde{W}_{d_2,\{1\}} \bigcup W^{1 \to \{2\}}_{d_2,\{1,3\}}, \tag{20}$$

where $W^{1 \to \{2\}}_{d_2,\{1,3\}} \subset \tilde{W}_{d_2,\{1,3\}}$, such that $|W^{1 \to \{2\}}_{d_2,\{1,3\}}| = u^{1 \to \{2\}}_{\{1,3\}} F$ bits. That is, the *assignment variable* $u^{j \to \mathcal{T}}_{\mathcal{S}} \in [0, a_{\mathcal{S}}]$ represents the fraction of the subfile $\tilde{W}_{\mathcal{S}}$ which is involved in the transmission from user j to the users in $\mathcal{T}$. Similarly, the unicast signal $X_{1 \to \{3\}}$ is given by

$$X_{1 \to \{3\}} = \tilde{W}_{d_3,\{1\}} \bigcup W^{1 \to \{3\}}_{d_3,\{1,2\}}, \tag{21}$$

where $W^{1 \to \{3\}}_{d_3,\{1,2\}} \subset \tilde{W}_{d_3,\{1,2\}}$, such that $|W^{1 \to \{3\}}_{d_3,\{1,2\}}| = u^{1 \to \{3\}}_{\{1,2\}} F$ bits.

The multicast signal $X_{1 \to \{2,3\}}$ is created by XORing the pieces $W^{1 \to \{2,3\}}_{d_2,\{1,3\}}$, and $W^{1 \to \{2,3\}}_{d_3,\{1,2\}}$, which are assumed to have equal size. That is, $X_{1 \to \{2,3\}}$ is defined by

$$X_{1 \to \{2,3\}} = W^{1 \to \{2,3\}}_{d_2,\{1,3\}} \oplus W^{1 \to \{2,3\}}_{d_3,\{1,2\}}, \tag{22}$$

where $W^{1 \to \{2,3\}}_{d_2,\{1,3\}} \subset \tilde{W}_{d_2,\{1,3\}}$ and $W^{1 \to \{2,3\}}_{d_3,\{1,2\}} \subset \tilde{W}_{d_3,\{1,2\}}$.

From (20)-(22), we observe that subfile $\tilde{W}_{d_2,\{1,3\}}$ contributes to both $X_{1 \to \{2\}}$, and $X_{1 \to \{2,3\}}$. Additionally, in the third transmission stage subfile $\tilde{W}_{d_2,\{1,3\}}$ contributes to both $X_{3 \to \{2\}}$, and $X_{3 \to \{1,2\}}$. Therefore, in order to ensure that $\tilde{W}_{d_2,\{1,3\}}$ is delivered to user 2, we have

$$W^{1 \to \{2\}}_{d_2,\{1,3\}} \bigcup W^{1 \to \{2,3\}}_{d_2,\{1,3\}} \bigcup W^{3 \to \{2\}}_{d_2,\{1,3\}} \bigcup W^{3 \to \{1,2\}}_{d_2,\{1,3\}} = \tilde{W}_{d_2,\{1,3\}}, \tag{23}$$

$$W^{1 \to \{2\}}_{d_2,\{1,3\}} \bigcap W^{1 \to \{2,3\}}_{d_2,\{1,3\}} \bigcap W^{3 \to \{2\}}_{d_2,\{1,3\}} \bigcap W^{3 \to \{1,2\}}_{d_2,\{1,3\}} = \phi. \tag{24}$$

2) Delivery Phase Constraints: Next, we describe the delivery phase in terms of linear constraints on the transmission variables $v_{j \to \mathcal{T}}$ and the assignment variables $u^{j \to \mathcal{T}}_{\mathcal{S}}$, which represent $|X_{j \to \mathcal{T}}|/F$ and $|W^{j \to \mathcal{T}}_{d_i,\mathcal{S}}|/F$, respectively.

First, the structure of the unicast signals in (20) and (21) is represented by

$$v_{1 \to \{2\}} = a_{\{1\}} + u^{1 \to \{2\}}_{\{1,3\}}, \quad v_{1 \to \{3\}} = a_{\{1\}} + u^{1 \to \{3\}}_{\{1,2\}}. \tag{25}$$

Similarly, for the second and third transmission stage, we have

$$v_{2 \to \{1\}} = a_{\{2\}} + u^{2 \to \{1\}}_{\{2,3\}}, \quad v_{2 \to \{3\}} = a_{\{2\}} + u^{2 \to \{3\}}_{\{1,2\}}, \tag{26}$$

$$v_{3 \to \{1\}} = a_{\{3\}} + u^{3 \to \{1\}}_{\{2,3\}}, \quad v_{3 \to \{2\}} = a_{\{3\}} + u^{3 \to \{2\}}_{\{1,3\}}. \tag{27}$$

The structure of the multicast signal in (22) is represented by

$$v_{1 \to \{2,3\}} = u^{1 \to \{2,3\}}_{\{1,3\}} = u^{1 \to \{2,3\}}_{\{1,2\}}. \tag{28}$$

Similarly, for the second and third transmission stage, we have

$$v_{2 \to \{1,3\}} = u^{1 \to \{2,3\}}_{\{2,3\}} = u^{1 \to \{2,3\}}_{\{1,2\}}, \tag{29}$$

$$v_{3 \to \{1,2\}} = u^{3 \to \{1,2\}}_{\{2,3\}} = u^{3 \to \{1,2\}}_{\{1,3\}}. \tag{30}$$

Additionally, (23) and (24) ensure the delivery of $\tilde{W}_{d_2,\{1,3\}}$ to user 2. Hence, we have

$$u^{1 \to \{2\}}_{\{1,3\}} + u^{1 \to \{2,3\}}_{\{1,3\}} + u^{3 \to \{2\}}_{\{1,3\}} + u^{3 \to \{1,2\}}_{\{1,3\}} = a_{\{1,3\}}. \tag{31}$$

Similarly, for subfiles $\tilde{W}_{d_3,\{1,2\}}$ and $\tilde{W}_{d_1,\{2,3\}}$, we have

$$u^{1 \to \{3\}}_{\{1,2\}} + u^{1 \to \{2,3\}}_{\{1,2\}} + u^{2 \to \{3\}}_{\{1,2\}} + u^{2 \to \{1,3\}}_{\{1,2\}} = a_{\{1,2\}}, \tag{32}$$

$$u^{2 \to \{1\}}_{\{2,3\}} + u^{2 \to \{1,3\}}_{\{2,3\}} + u^{3 \to \{1\}}_{\{2,3\}} + u^{3 \to \{1,2\}}_{\{2,3\}} = a_{\{2,3\}}. \tag{33}$$

Therefore, the set of feasible linear delivery schemes for a three-user system is defined by (25)-(33), and $u^{j \to \mathcal{T}}_{\mathcal{S}} \in [0, a_{\mathcal{S}}]$.

C. Delivery Phase: K-User System

In general, the unicast signal transmitted by user j to user i is defined by

$$X_{j \to \{i\}} = \tilde{W}_{d_i,\{j\}} \bigcup \Bigg(\bigcup_{\mathcal{S} \subset [K] \setminus \{i\} \,:\, j \in \mathcal{S}, |\mathcal{S}| \ge 2} W^{j \to \{i\}}_{d_i,\mathcal{S}} \Bigg), \tag{34}$$

where $W^{j \to \{i\}}_{d_i,\mathcal{S}} \subset \tilde{W}_{d_i,\mathcal{S}}$ such that $|W^{j \to \{i\}}_{d_i,\mathcal{S}}| = u^{j \to \{i\}}_{\mathcal{S}} F$ bits. While, user j constructs the multicast signal $X_{j \to \mathcal{T}}$, such that the piece intended for user $i \in \mathcal{T}$, which we denote by $W^{j \to \mathcal{T}}_{d_i}$, is stored at users $\{j\} \cup (\mathcal{T} \setminus \{i\})$. That is, $X_{j \to \mathcal{T}}$ is constructed using the side information at the sets

$$\mathcal{B}^{j \to \mathcal{T}}_i \triangleq \Big\{ \mathcal{S} \subset [K] \setminus \{i\} : \{j\} \cup (\mathcal{T} \setminus \{i\}) \subset \mathcal{S} \Big\}, \tag{35}$$

which represents the subfiles stored at users $\{j\} \cup (\mathcal{T} \setminus \{i\})$ and not available at user $i \in \mathcal{T}$. In turn, we have

$$X_{j \to \mathcal{T}} = \oplus_{i \in \mathcal{T}} W^{j \to \mathcal{T}}_{d_i} = \oplus_{i \in \mathcal{T}} \Bigg(\bigcup_{\mathcal{S} \in \mathcal{B}^{j \to \mathcal{T}}_i} W^{j \to \mathcal{T}}_{d_i,\mathcal{S}} \Bigg). \tag{36}$$

Algorithm 1 D2D Delivery Procedure

Input: d, a, u, v, and $\tilde{W}_{n,\mathcal{S}}$
Output: $X_{j\to\mathcal{T}}$, $\forall j \in [K]$, $\forall \mathcal{T} \subsetneq_\phi [K]\setminus\{j\}$
  # Partitioning
for $\{\mathcal{S} \subset [K] : 2 \le |\mathcal{S}| \le K-1\}$ **do**
  for $\{i \in [K] : i \notin \mathcal{S}\}$ **do**
    Divide $\tilde{W}_{d_i,\mathcal{S}}$ into $W^{j\to\mathcal{T}}_{d_i,\mathcal{S}}$, $\forall j \in \mathcal{S}, \forall \mathcal{T} \subset \{i\} \cup (\mathcal{S} \setminus \{j\})$ s.t. $i \in \mathcal{T}$, such that $|W^{j\to\mathcal{T}}_{d_i,\mathcal{S}}| = u^{j\to\mathcal{T}}_{\mathcal{S}} F$ bits.
  end for
end for
  # Transmission stage j
for $j \in [K]$ **do**
  for $\mathcal{T} \subsetneq_\phi [K]\setminus\{j\}$ **do**
    if $\mathcal{T} = \{i\}$ **then**
      $$X_{j\to\{i\}} \leftarrow \tilde{W}_{d_i,\{j\}} \bigcup \Bigg(\bigcup_{\mathcal{S}\subset[K]\setminus\{i\}\ j\in\mathcal{S},|\mathcal{S}|\ge 2} W^{j\to\{i\}}_{d_i,\mathcal{S}} \Bigg)$$
    else
      $$X_{j\to\mathcal{T}} \leftarrow \oplus_{i\in\mathcal{T}} \Bigg(\bigcup_{\mathcal{S}\in\mathcal{B}^{j\to\mathcal{T}}_i} W^{j\to\mathcal{T}}_{d_i,\mathcal{S}} \Bigg)$$
    end if
  end for
end for

Remark 2: The definition of the multicast signals in (36) allows flexible utilization of the side-information, i.e., $X_{j\to\mathcal{T}}$ is not defined only in terms of the side-information stored exclusively at users $\{j\} \cup (\mathcal{T} \setminus \{i\})$ as in [6]. Furthermore, a delivery scheme with the multicast signals $X_{j\to\mathcal{T}} = \oplus_{i\in\mathcal{T}} W^{j\to\mathcal{T}}_{d_i,\{j\}\cup(\mathcal{T}\setminus\{i\})}$ is suboptimal in general.

The set of feasible linear delivery schemes, $\mathfrak{D}(a)$, is defined by

$$v_{j\to\{i\}} = a_{\{j\}} + \sum_{\mathcal{S}\subset[K]\setminus\{i\}: j\in\mathcal{S}, |\mathcal{S}|\ge 2} u^{j\to\{i\}}_{\mathcal{S}}, \quad \forall j \in [K],\ \forall i \in \mathcal{T}, \tag{37}$$

$$v_{j\to\mathcal{T}} = \sum_{\mathcal{S}\in\mathcal{B}^{j\to\mathcal{T}}_i} u^{j\to\mathcal{T}}_{\mathcal{S}}, \quad \forall j\in[K],\ \forall\, \mathcal{T} \subsetneq_\phi [K]\setminus\{j\},\ \forall i\in\mathcal{T}, \tag{38}$$

$$\sum_{j\in\mathcal{S}} \sum_{\mathcal{T}\subset\{i\}\cup(\mathcal{S}\setminus\{j\})\ :\ i\in\mathcal{T}} u^{j\to\mathcal{T}}_{\mathcal{S}} = a_{\mathcal{S}}, \quad \forall\, i \notin \mathcal{S}, \quad \forall\, \mathcal{S} \subset [K] \quad \text{s.t.} \quad 2 \le |\mathcal{S}| \le K-1, \tag{39}$$

$$0 \le u^{j\to\mathcal{T}}_{\mathcal{S}} \le a_{\mathcal{S}}, \quad \forall j \in [K],\ \forall\, \mathcal{T} \subsetneq_\phi [K]\setminus\{j\},\ \forall\, \mathcal{S}\in\mathcal{B}^{j\to\mathcal{T}}, \tag{40}$$

where $\mathcal{B}^{j\to\mathcal{T}} \triangleq \bigcup_{i\in\mathcal{T}} \mathcal{B}^{j\to\mathcal{T}}_i$. Note that (37) follows from the structure of the unicast signals in (34), (38) follows from the structure of the multicast signals in (36), (39) generalizes the constraints in (31)-(33). The delivery procedure is summarized in Algorithm 1.

Next example shows the suboptimality of delivery schemes that do not allow flexible utilization of the side-information, as pointed out in Remark 2. By contrast, our delivery scheme achieves the delivery load memory trade-off with uncoded placement, $R^*_{\mathfrak{A}}(m)$.

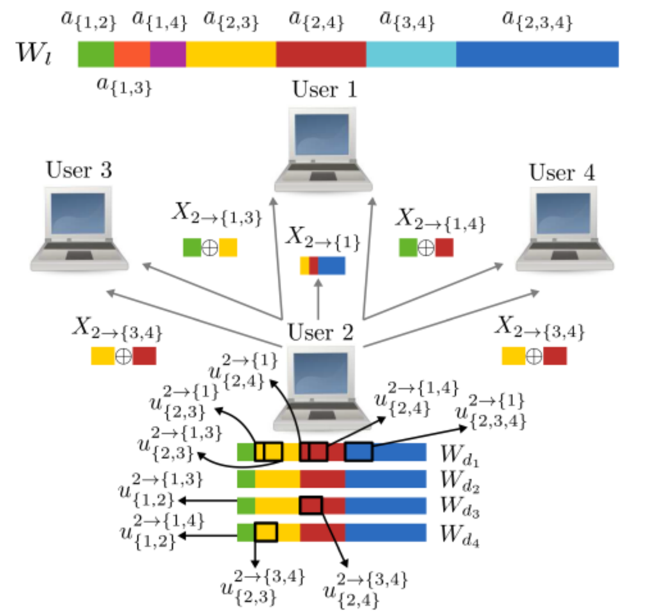


Fig. 3. Example $K = N = 4$, and $\boldsymbol{m} = [0.2, 0.7, 0.7, 0.7]$.

Example 2: For $K = N = 4$ and $m = [0.2, 0.7, 0.7, 0.7]$, we have $R^*_{\mathfrak{A},\mathfrak{D}}(m) = R^*_{\mathfrak{A}}(m) = 1.05$, and the optimal caching scheme is as follows:

Placement Phase: Each file W_n is divided into seven subfiles, such that $a_{\{1,2\}} = a_{\{1,3\}} = a_{\{1,4\}} = 0.2/3$, $a_{\{2,3\}} = a_{\{2,4\}} = a_{\{3,4\}} = 0.5/3$, and $a_{\{2,3,4\}} = 0.3$.

Delivery Phase: We have the D2D transmissions $X_{2\to\{1\}}$, $X_{2\to\{1,3\}}$, $X_{2\to\{1,4\}}$, $X_{2\to\{3,4\}}$, $X_{3\to\{1\}}$, $X_{3\to\{1,2\}}$, $X_{3\to\{1,4\}}$, $X_{3\to\{2,4\}}$, $X_{4\to\{1\}}$, $X_{4\to\{1,2\}}$, $X_{4\to\{1,3\}}$, and $X_{4\to\{2,3\}}$. In particular, we have $v_{2\to\{1\}} = v_{3\to\{1\}} = v_{4\to\{1\}} = 0.4/3$, $v_{2\to\{1,3\}} = v_{2\to\{1,4\}} = v_{3\to\{1,2\}} = v_{3\to\{1,4\}} = v_{4\to\{1,2\}} = v_{4\to\{1,3\}} = 0.2/3$, and $v_{2\to\{3,4\}} = v_{3\to\{2,4\}} = v_{4\to\{2,3\}} = 0.25/3$. More specifically, the signals transmitted by user 2 are defined as follows

- $|X_{2\to\{1\}}|/F = v_{2\to\{1\}} = u^{2\to\{1\}}_{\{2,3\}} + u^{2\to\{1\}}_{\{2,4\}} + u^{2\to\{1\}}_{\{2,3,4\}} = (0.05 + 0.05 + 0.3)/3$.
- $|X_{2\to\{1,3\}}|\,/\,F = v_{2\to\{1,3\}} = u^{2\to\{1,3\}}_{\{1,2\}} = u^{2\to\{1,3\}}_{\{2,3\}} = 0.2/3$.
- $|X_{2\to\{1,4\}}|\,/\,F = v_{2\to\{1,4\}} = u^{2\to\{1,4\}}_{\{1,2\}} = u^{2\to\{1,4\}}_{\{2,4\}} = 0.2/3$.
- $|X_{2\to\{3,4\}}|\,/\,F = v_{2\to\{3,4\}} = u^{2\to\{3,4\}}_{\{2,3\}} = u^{2\to\{3,4\}}_{\{2,4\}} = 0.25/3$.

Note that the signals transmitted by users 3 and 4 have similar structure to the signals transmitted by user 2, which are illustrated in Fig. 3. If we restrict the design of the D2D signals to be in the form of $X_{j\to\mathcal{T}} = \oplus_{i\in\mathcal{T}} W^{j\to\mathcal{T}}_{d_i,\{j\}\cup(\mathcal{T}\setminus\{i\})}$, i.e., without the flexibility in utilizing the side information, we achieve a delivery load equal to 1.6 compared with the optimal load $R^*_{\mathfrak{A}}(m) = 1.05$.

V. Caching Scheme: Achievability

Next, we explicitly define the caching schemes that achieve the delivery loads defined in Theorems 4, 5, and 6.

A. Achievability Proof of Theorem 4

Next, we explain how the caching scheme in [6] can be tailored to systems with unequal cache sizes. Recall that for a homogeneous system where $m_k = m$, $\forall k$, in the placement phase, W_n is divided into subfiles $\tilde{W}_{n,\mathcal{S}}, \mathcal{S} \subset [K]$, where $|\mathcal{S}| \in \{t, t+1\}$ for $t \leq \sum_{k=1}^K m_k \leq t+1$ and $t \in [K-1]$ [6]. More specifically, subfiles stored at the same number of users have equal size, i.e., $|\tilde{W}_{n,\mathcal{S}}| = |\tilde{W}_{n,\mathcal{S}'}|$ if $|\mathcal{S}| = |\mathcal{S}'|$. In order to accommodate the heterogeneity in cache sizes, we generalize the placement scheme in [6], by allowing subfiles stored at the same number of users to have different sizes. The delivery procedure in [6] is generalized as follows. First, we further divide $\tilde{W}_{d_i,\mathcal{S}}$ into $|\mathcal{S}|$ pieces, $W_{d_i,\mathcal{S}}^{j\to\mathcal{S}\setminus\{j\}\cup\{i\}}$, $j \in \mathcal{S}$, such that

$$\left|W_{d_i,\mathcal{S}}^{j\to\mathcal{S}\setminus\{j\}\cup\{i\}}\right| = \begin{cases} \eta_j F, & \text{if } |\mathcal{S}| = t. \\ \theta_j F, & \text{if } |\mathcal{S}| = t+1. \end{cases} \tag{41}$$

The multicast signal $X_{j\to\mathcal{T}}$ is constructed such that the piece requested by user i is cached by the remaining $\mathcal{T} \setminus \{i\}$ users. That is, user j transmits the signals $X_{j\to\mathcal{T}} = \oplus_{i\in\mathcal{T}} W_{d_i,\{j\}\cup\mathcal{T}\setminus\{i\}}^{j\to\mathcal{T}}$, $\forall \mathcal{T} \subset [K] \setminus \{j\}$ and $|\mathcal{T}| \in \{t, t+1\}$. For example, for $K = 4$ and $t = 2$, we have

$$X_{j\to\{i_1,i_2\}} = W_{d_{i_1},\{j,i_2\}}^{j\to\{i_1,i_2\}} \oplus W_{d_{i_2},\{j,i_1\}}^{j\to\{i_1,i_2\}}, \tag{42}$$

$$X_{j\to\{i_1,i_2,i_3\}} = W_{d_{i_1},\{j,i_2,i_3\}}^{j\to\{i_1,i_2,i_3\}} \oplus W_{d_{i_2},\{j,i_1,i_3\}}^{j\to\{i_1,i_2,i_3\}} \oplus W_{d_{i_3},\{j,i_1,i_2\}}^{j\to\{i_1,i_2,i_3\}}. \tag{43}$$

In turn, the D2D delivery load is given as

$$R^*_{\mathfrak{A},\mathfrak{D}}(\boldsymbol{m}) = \binom{K-1}{t} \sum_{j=1}^K \eta_j + \binom{K-1}{t+1} \sum_{j=1}^K \theta_j. \tag{44}$$

Next, we need to choose η_j and θ_j taking into account the feasibility of the placement phase. To do so, we need to choose a non-negative solution to the following equations

$$\binom{K-1}{t-1}\eta_k + \binom{K-2}{t-2} \sum_{i\in[K]\setminus\{k\}} \eta_i + \binom{K-1}{t}\theta_k + \binom{K-2}{t-1} \sum_{i\in[K]\setminus\{k\}} \theta_i = m_k, \quad \forall k \in [K], \tag{45}$$

$$\binom{K-1}{t-1}\sum_{i\in[K]} \eta_i + \binom{K-1}{t} \sum_{i\in[K]} \theta_i = 1, \tag{46}$$

which can be simplified to

$$\sum_{i=1}^K \eta_i = \frac{t+1-\sum_{i=1}^K m_k}{\binom{K-1}{t-1}}, \tag{47}$$

$$\eta_k + \frac{K-t-1}{t}\theta_k = \frac{1+(K-2)m_k - \sum\limits_{i\in[K]\setminus\{k\}} m_i}{(K-t)\binom{K-1}{t-1}}, \quad \forall k, \tag{48}$$

By combining (44), (47), and (48), one can show that the D2D delivery load is given as

$$R^*_{\mathfrak{A},\mathfrak{D}}(\boldsymbol{m}) = \binom{K-t}{t}\left(t+1-\sum_{k=1}^K m_k\right) + \binom{K-t-1}{t+1}\left(\sum_{k=1}^K m_k - t\right). \tag{49}$$

Observe that there always exists a non-negative solution to (47) and (48), since we have $(K-2)\, m_1 \geq \sum_{k=2}^K m_k - 1$. For instance, one can assume that $\binom{K-1}{t-1}\eta_k = \rho_k\left(t+1-\sum_{i=1}^K m_i\right)$, where $\sum_{k=1}^K \rho_k = 1$ and $0 \leq \rho_k \leq \dfrac{1+(K-2)m_k - \sum_{i\in[K]\setminus\{k\}} m_i}{(K-t)(t+1-\sum_{i=1}^K m_k)}$, which guarantee that $\eta_k, \theta_k \geq 0$.

Remark 3: For nodes with equal cache sizes, the proposed scheme reduces to the scheme proposed in [6]. In particular, for $m_k = t/K, \forall k$, we get $\theta_j = 0, \forall j$ and $\eta_j = 1/\left(t\binom{K}{t}\right), \forall j$.

B. Achievability Proof of Theorem 5

For $(K-l-1)m_l < \sum_{i=l+1}^K m_i - 1$ and $(K-l-2)m_{l+1} \geq \sum_{i=l+2}^K m_i - 1$, where $l \in [K-2]$, in the placement phase, each file W_n is partitioned into subfiles $\tilde{W}_{n,\{i\}}, i \in \{l+1,\dots,K\}$, $\tilde{W}_{n,\{j,i\}}, j \in [l], i \in \{l+1,\dots,K\}$, and $\tilde{W}_{n,\mathcal{S}}, \mathcal{S} \subset \{l+1,\dots,K\}, |\mathcal{S}| = 2$, which satisfy

$$\sum_{j=l+1}^K a_{\{j\}} = 2 - \sum_{k=1}^K m_k, \tag{50a}$$

$$\sum_{\mathcal{S}\subset\{l+1,\dots,K\}:|\mathcal{S}|=2} a_{\mathcal{S}} = \sum_{i=l+1}^K m_i - 1, \tag{50b}$$

$$\sum_{j=l+1}^K a_{\{i,j\}} = m_j, \quad i \in [l], \tag{50c}$$

$$a_{\{j\}} + \sum_{\mathcal{S}\subset[K]:|\mathcal{S}|=2, j\in\mathcal{S}} a_{\mathcal{S}} = m_j, \quad j = l+1,\dots,K. \tag{50d}$$

In particular, we choose any non-negative solution to (50) that satisfies

1) For $j \in \{l+1,\dots,K\}$, $a_{\{i_1,j\}} \leq a_{\{i_2,j\}}$ if $i_1 < i_2$, which is feasible because $m_{i_1} \leq m_{i_2}$.
2) For $\{i,j\} \subset \{l+1,\dots,K\}$, $a_{\{l,i\}} + a_{\{l,j\}} \leq a_{\{i,j\}}$, which is also feasible because $(K-l-1)m_l < \sum_{i=l+1}^K m_i - 1$.

In the delivery phase, we have the following multicast transmissions:

- Multicast to user 1: For $j \in \{l+1,\dots,K\}$ and $i \in [K] \setminus \{1,j\}$, we choose $v_{j\to\{1,i\}} = a_{\{1,j\}}$.

$$\sum_{j=l+1}^K \sum_{i\in[K]\setminus\{1,j\}} v_{j\to\{1,i\}} = \sum_{j=l+1}^K \sum_{i\in[K]\setminus\{1,j\}} a_{\{1,j\}} = (K-2)m_1. \tag{51}$$

- Multicast to user 2: For $j \in \{l+1,\ldots,K\}$ and $i \in [K]\setminus\{1,2,j\}$, we choose $v_{j\to\{2,i\}} = a_{\{2,j\}}$.

$$\sum_{j=l+1}^{K} v_{j\to\{1,2\}} + \sum_{j=l+1}^{K}\sum_{i\in[K]\setminus\{1,2,j\}} v_{j\to\{2,i\}} = \sum_{j=l+1}^{K} a_{\{1,j\}} + \sum_{j=l+1}^{K}\sum_{i\in[K]\setminus\{1,2,j\}} a_{\{2,j\}} = m_1 + (K-3)m_2. \quad (52)$$

- Multicast to user $k \in \{3,\ldots,l\}$: Similarly, we have

$$\sum_{j=l+1}^{K} v_{j\to\{1,l\}} + \cdots + \sum_{j=l+1}^{K} v_{j\to\{k-1,l\}} + \sum_{j=l+1}^{K}\sum_{i\in[K]\setminus\{1,\ldots,k,j\}} v_{j\to\{l,i\}} = \sum_{j=l+1}^{K} a_{\{1,j\}} + \ldots + \sum_{j=l+1}^{K} a_{\{k-1,j\}} + \sum_{j=l+1}^{K}\sum_{i\in[K]\setminus\{1,\ldots,k,j\}} a_{\{l,j\}} = \sum_{i=1}^{k-1} m_i + (K-k-1)m_l. \quad (53)$$

- Multicast to users $\{l+1,\ldots,K\}$: For $\{i_1,i_2\}\subset\{l+1,\ldots,K\}$, we have $a_{\{i_1,i_2\}} = v_{i_1\to\{i_2,j\}} + v_{i_2\to\{i_1,j\}}$, $\forall j \in \{l{+}1,\ldots,K\}\setminus\{i_1,i_2\}$, i.e., we have $(K{-}l{-}2)\binom{K-l}{2}$ equations in $(K{-}l{-}2)\binom{K-l}{2}$ unknowns. In turn, we have

$$\sum_{j=l+1}^{K}\sum_{\mathcal{S}\subset\{l+1,\ldots,K\}\setminus\{j\}:|\mathcal{S}|=2} v_{j\to\mathcal{S}} = \left(\frac{K-l-2}{2}\right)\sum_{\mathcal{S}\subset\{l+1,\ldots,K\}:|\mathcal{S}|=2} a_{\mathcal{S}} = \left(\frac{K-l-2}{2}\right)\left(\sum_{i=l+1}^{K} m_i - 1\right). \quad (54)$$

Therefore, the delivery load due to multicast transmissions is given by

$$\sum_{j=l+1}^{K}\sum_{\mathcal{S}\subset[K]\setminus\{j\}:|\mathcal{S}|=2} v_{j\to\mathcal{S}} = \sum_{j=l+1}^{K}\left(\sum_{i\in[K]\setminus\{1,j\}} v_{j\to\{1,i\}} + \cdots + \sum_{i\in[K]\setminus\{1,\ldots,l,j\}} v_{j\to\{l,i\}} + \sum_{\mathcal{S}\subset\{l+1,\ldots,K\}\setminus\{j\}:|\mathcal{S}|=2} v_{j\to\mathcal{S}}\right) = \sum_{i=1}^{l}(K-i-1)m_i + \left(\frac{K-l-2}{2}\right)\left(\sum_{i=l+1}^{K} m_i - 1\right). \quad (55)$$

We also need the following unicast transmissions.

- Unicast to user 1:

$$\sum_{j=l+1}^{K} v_{j\to\{1\}} = \sum_{j=l+1}^{K} a_{\{j\}} + \sum_{i=2}^{l}\sum_{j=l+1}^{K}(a_{\{i,j\}} - a_{\{1,j\}}) + \sum_{\{i,j\}\subset\{l+1,\ldots,K\}}(a_{\{i,j\}} - a_{\{1,i\}} - a_{\{1,j\}}) = \left(2 - \sum_{k=1}^{K} m_k\right) + \sum_{i=2}^{l} m_i + \left(\sum_{i=l+1}^{K} m_i - 1\right) - (K-2)m_1 = 1-(K-1)m_1. \quad (56)$$

- Unicast to user 2:

$$\sum_{j=l+1}^{K} v_{j\to\{2\}} = \sum_{j=l+1}^{K} a_{\{j\}} + \sum_{i=3}^{l}\sum_{j=l+1}^{K}(a_{\{i,j\}} - a_{\{2,j\}}) + \sum_{\{i,j\}\subset\{l+1,\ldots,K\}}(a_{\{i,j\}} - a_{\{2,i\}} - a_{\{2,j\}}) = 1-(K-2)m_2 - m_1. \quad (57)$$

- Unicast to user $k \in \{3,\ldots,l\}$: Similarly, we have

$$\sum_{j=l+1}^{K} v_{j\to\{l\}} = 1-(K-k)m_k - m_{k-1} - \ldots - m_1. \quad (58)$$

- Unicast to users $\{l+1,\ldots,K\}$:

$$\sum_{j=l+1}^{K}\sum_{i=l+1,i\neq j}^{K} v_{j\to\{i\}} = (K-l-1)\sum_{j=l+1}^{K} a_{\{j\}} = (K-l-1)\left(2 - \sum_{k=1}^{K} m_k\right). \quad (59)$$

Therefore, the delivery load due to unicast transmissions is given by

$$\sum_{j=l+1}^{K}\sum_{i=1,i\neq j}^{K} v_{j\to\{i\}} = l - \sum_{i=1}^{l}(K+l-2i)m_i + (K-l-1)\left(2 - \sum_{k=1}^{K} m_k\right). \quad (60)$$

By adding (55) and (60), we get the total D2D delivery load given by (15).

C. Achievability Proof of Theorem 6

For $\sum_{i=1}^{K} m_i \geq K{-}1$, in the placement phase, each file W_n is partitioned into subfiles $\tilde{W}_{n,[K]\setminus\{i\}}, i \in [K]$ and $\tilde{W}_{n,[K]}$, such that

$$a_{[K]} = \sum_{i=1}^{K} m_i - (K-1), \quad a_{[K]\setminus\{k\}} = 1 - m_k, \quad k \in [K]. \quad (61a)$$

In the delivery phase, for $(K-l-1)m_l < \sum_{i=l+1}^{K} m_i - 1$ and $(K-l-2)m_{l+1} \geq \sum_{i=l+2}^{K} m_i - 1$, where $l \in [K-2]$, we have the following transmissions

$$X_{K\to[i]} = \oplus_{k\in[i]} W^{K\to[i]}_{d_k,[K]\setminus\{k\}}, \quad i \in [l], \tag{62}$$

$$X_{j\to[K]\setminus\{j\}} = \oplus_{k\in[K]\setminus\{j\}} W^{j\to[K]\setminus\{j\}}_{d_k,[K]\setminus\{k\}}, \quad j\in\{l+1,\dots,K\}. \tag{63}$$

In particular, we have

$$v_{K\to[i]} = u^{K\to[i]}_{[K]\setminus\{k\}} = m_{i+1} - m_i, i \in [l-1], k \in [i], \tag{64}$$

$$v_{K\to[l]} = u^{K\to[l]}_{[K]\setminus\{k\}} = \frac{\sum_{j=l+1}^{K} m_j - 1 - (K-l-1)m_l}{K-l-1}, \quad k \in [l], \tag{65}$$

$$v_{j\to[K]\setminus\{j\}} = u^{j\to[K]\setminus\{j\}}_{[K]\setminus\{k\}} = \frac{(K-l-1)m_j + 1 - \sum_{i=l+1}^{K} m_i}{K-l-1}, \quad j\in\{l+1,\dots,K\}, \quad k\in[K]\setminus\{j\}. \tag{66}$$

Therefore, the D2D delivery load is given by

$$R^*_{\mathfrak{A},\mathfrak{D}}(m) = v_{K\to[l]} + \sum_{i=1}^{l} v_{K\to[i]} + \sum_{j=l+1}^{K} v_{j\to[K]\setminus\{j\}}, \tag{67}$$

$$= 1 - m_1. \tag{68}$$

VI. Optimality With Uncoded Placement and One-Shot Delivery

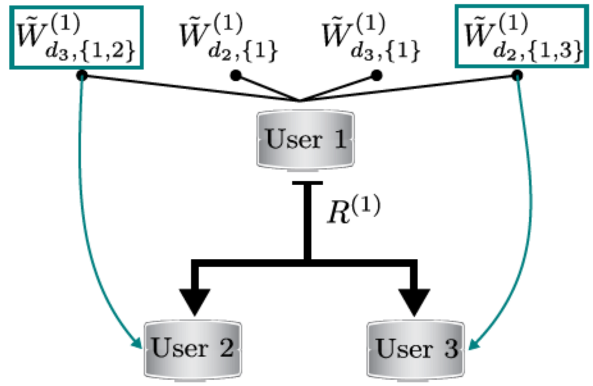
In this section, we first prove the lower bound in Theorem 2. Then, we present the converse proofs for Theorems 3, 4, and 5.

A. Proof of Theorem 2

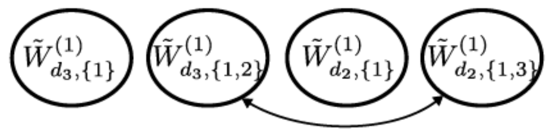
First, we show that the D2D-based delivery assuming uncoded placement and one-shot delivery can be represented by K index-coding problems, i.e., each D2D transmission stage is equivalent to an index-coding problem. In particular, for any allocation $\boldsymbol{a} \in \mathfrak{A}(\boldsymbol{m})$, we assume that each subfile $\tilde{W}_{d_i,\mathcal{S}}$ consists of $|\mathcal{S}|$ disjoint pieces $\tilde{W}^{(j)}_{d_i,\mathcal{S}}$, $j \in \mathcal{S}$, where $|\tilde{W}^{(j)}_{d_i,\mathcal{S}}| = a^{(j)}_{\mathcal{S}} F$ bits, i.e., $a_{\mathcal{S}} = \sum_{j\in\mathcal{S}} a^{(j)}_{\mathcal{S}}$. Additionally, the file pieces with superscript (j) represent the messages in the jth index-coding problem.

For instance, consider the first index-coding problem in a three-user system, in which user 1 acts as a server, see Fig. 4(a). User 1 needs to deliver $\tilde{W}^{(1)}_{d_2,\{1\}}, \tilde{W}^{(1)}_{d_2,\{1,3\}}$ to user 2, and $\tilde{W}^{(1)}_{d_3,\{1\}}, \tilde{W}^{(1)}_{d_3,\{1,2\}}$ to user 3. User 2 has access to $\tilde{W}^{(1)}_{d_3,\{1,2\}}$, and user 3 has access to $\tilde{W}^{(1)}_{d_2,\{1,3\}}$. The index coding problem depicted in Fig. 4(a) can be represented by the directed graph shown in Fig. 4(b), where the nodes represent the messages and a directed edge from $\tilde{W}^{(1)}_{*,\mathcal{S}}$ to $\tilde{W}^{(1)}_{d_i,*}$ exists if $i \in \mathcal{S}$ [8]. Furthermore, by applying the acyclic index-coding bound [16, Corollary 1] on Fig. 4(b), we get

$$R^{(1)}F \geq \sum_{i=1}^{K-1} \sum_{\mathcal{S}\subset[K]:1\in\mathcal{S},\{q_1,\dots,q_i\}\cap\mathcal{S}=\phi} |\tilde{W}^{(1)}_{d_{q_i},\mathcal{S}}|, \tag{69}$$



(a) Index-coding setup.



(b) Graph representation.

Fig. 4. Index-coding problem for $K = 3$, and $j = 1$.

where $\boldsymbol{q} \in \mathcal{P}_{\{2,3\}}$ [8], [9]. In particular, for $K=3$, we have

$$R^{(1)}F \geq |\tilde{W}^{(1)}_{d_2,\{1\}}| + |\tilde{W}^{(1)}_{d_3,\{1\}}| + |\tilde{W}^{(1)}_{d_2,\{1,3\}}|, \quad \boldsymbol{q}=[2,3], \tag{70}$$

$$R^{(1)}F \geq |\tilde{W}^{(1)}_{d_2,\{1\}}| + |\tilde{W}^{(1)}_{d_3,\{1\}}| + |\tilde{W}^{(1)}_{d_3,\{1,2\}}|, \quad \boldsymbol{q}=[3,2]. \tag{71}$$

Hence, for a given partitioning $a^{(j)}_{\mathcal{S}}$, by taking the convex combination of (70), and (71), we get

$$R^{(1)}(a^{(1)}_{\mathcal{S}}, \alpha_{\boldsymbol{q}}) \geq 2a^{(1)}_{\{1\}} + \alpha_{[2,3]} a^{(1)}_{\{1,3\}} + \alpha_{[3,2]} a^{(1)}_{\{1,2\}}, \tag{72}$$

where $\alpha_{\boldsymbol{q}} \geq 0$, and $\alpha_{[2,3]} + \alpha_{[3,2]} = 1$. Similarly, we have

$$R^{(2)}(a^{(2)}_{\mathcal{S}}, \alpha_{\boldsymbol{q}}) \geq 2a^{(2)}_{\{2\}} + \alpha_{[1,3]} a^{(2)}_{\{2,3\}} + \alpha_{[3,1]} a^{(2)}_{\{1,2\}}, \tag{73}$$

$$R^{(3)}(a^{(3)}_{\mathcal{S}}, \alpha_{\boldsymbol{q}}) \geq 2a^{(3)}_{\{3\}} + \alpha_{[1,2]} a^{(3)}_{\{2,3\}} + \alpha_{[2,1]} a^{(3)}_{\{1,3\}}. \tag{74}$$

Hence, for given $a^{(j)}_{\mathcal{S}}$ and $\alpha_{\boldsymbol{q}}$, the D2D delivery load $\sum_{j=1}^{3} R^{(j)}(a^{(j)}_{\mathcal{S}}, \alpha_{\boldsymbol{q}})$ is lower bounded by the sum of the right-hand side of (72)-(74). Furthermore, for K-user systems, $R^{(j)}(a^{(j)}_{\mathcal{S}}, \alpha_{\boldsymbol{q}})$ is lower bounded by

$$\tilde{R}^{(j)}(a^{(j)}_{\mathcal{S}}, \alpha_{\boldsymbol{q}}) \triangleq (K-1)\, a^{(j)}_{\{j\}} + \sum_{\substack{\mathcal{S}\subset[K]\,:\, j\in\mathcal{S},\\ 2\leq|\mathcal{S}|\leq K-1}} \left(\sum_{i=1}^{K-|\mathcal{S}|} \sum_{\substack{\boldsymbol{q}\in\mathcal{P}_{[K]\setminus\{j\}}:\ q_{i+1}\in\mathcal{S},\\ \{q_1,\dots,q_i\}\cap\mathcal{S}=\phi}} i\ \alpha_{\boldsymbol{q}} \right) a^{(j)}_{\mathcal{S}}. \tag{75}$$

By taking the minimum over all feasible allocations and partitions, we get

$$R^*_{\mathfrak{A}}(\alpha_{\boldsymbol{q}}) \geq \min_{a^{(j)}_{\mathcal{S}}\geq 0} \sum_{j=1}^{K} \tilde{R}^{(j)}(a^{(j)}_{\mathcal{S}}, \alpha_{\boldsymbol{q}}) \tag{76a}$$

$$\text{subject to} \sum_{\mathcal{S}\subsetneq_{\phi}[K]} \sum_{j\in\mathcal{S}} a^{(j)}_{\mathcal{S}} = 1, \tag{76b}$$

$$\sum_{\mathcal{S}\subset[K]:k\in\mathcal{S}} \sum_{j\in\mathcal{S}} a^{(j)}_{\mathcal{S}} \leq m_k, \quad \forall\, k\in[K]. \tag{76c}$$

The dual of the linear program in (76) is given by

$$\max_{\lambda_0\in\mathbb{R},\lambda_k\geq 0} \ -\lambda_0-\sum_{k=1}^{K} m_k\lambda_k \tag{77a}$$

$$\text{subject to } \lambda_0+\sum_{k\in\mathcal{S}}\lambda_k+\gamma_{\mathcal{S}}\geq 0, \quad \forall\, \mathcal{S}\subsetneq_{\phi}[K], \tag{77b}$$

where $\gamma_{\mathcal{S}}$ is defined in (12), λ_0, and λ_k are the dual variables associated with (76b), and (76c), respectively. Finally, by taking the maximum over all possible convex combinations $\alpha_{\boldsymbol{q}}, \forall \boldsymbol{q} \in \mathcal{P}_{[K]\setminus\{j\}}, \forall j \in [K]$, we get the lower bound in Theorem 2.

B. Converse Proof of Theorem 3

Next, we simplify the bound in Theorem 2 by averaging over all permutations $\boldsymbol{q} \in \mathcal{P}_{[K]\setminus\{j\}}$. In particular, by substituting $\alpha_{\boldsymbol{q}}=1/(K-1)!$ in Theorem 2, for $2\leq|\mathcal{S}|\leq K-1$ we get

$$\gamma_{\mathcal{S}}=\min_{j\in\mathcal{S}}\Bigg\{\sum_{i=1}^{K-|\mathcal{S}|}\ \sum_{\substack{\boldsymbol{q}\in\mathcal{P}_{[K]\setminus\{j\}}:\, q_{i+1}\in\mathcal{S},\\ \{q_1,\ldots,q_i\}\cap\mathcal{S}=\phi}} i/(K-1)!\Bigg\}, \tag{78}$$

$$=\sum_{i=1}^{K-|\mathcal{S}|}\frac{i}{(K-1)!}\binom{K-|\mathcal{S}|}{i}\, i!\,(|\mathcal{S}|-1)\,(K-i-2)!, \tag{79}$$

$$=\frac{(K-|\mathcal{S}|)!\,(|\mathcal{S}|-1)!}{(K-1)!}\sum_{i=1}^{K-|\mathcal{S}|} i\binom{K-i-2}{|\mathcal{S}|-2}, \tag{80}$$

$$=\frac{(K-|\mathcal{S}|)!\,(|\mathcal{S}|-1)!}{(K-1)!}\binom{K-1}{|\mathcal{S}|}=\frac{K-|\mathcal{S}|}{|\mathcal{S}|}, \tag{81}$$

where (79) follows from the number of vectors $\boldsymbol{q}\in\mathcal{P}_{[K]\setminus\{j\}}$ such that $q_{i+1}\in\mathcal{S}$, and $\{q_1,\ldots,q_i\}\cap\mathcal{S}=\phi$. In particular, for given $j\in[K]$, $\mathcal{S}\subset[K]$ such that $j\in\mathcal{S}$, and $i\in\{1,\ldots,K-|\mathcal{S}|\}$, there are $\binom{K-|\mathcal{S}|}{i}$ $i!$ choices for $\{q_1,\ldots,q_i\}$, $(|\mathcal{S}|-1)$ choices for q_{i+1}, and $(K-i-2)!$ choices for the remaining elements in $[K]\setminus(\{j\}\cup\{q_1,\ldots,q_{i+1}\})$. In turn, for $m_k=m$, $\forall k\in[K]$ and $|\mathcal{S}|=l$, the lower bound in Theorem 2 simplifies to

$$R^*_{\mathfrak{A}}(\boldsymbol{m})\geq\max_{\lambda_0\in\mathbb{R},\lambda\geq 0} \ -\lambda_0-Km\lambda \tag{82a}$$

$$\text{subject to } \lambda_0+l\lambda+\frac{K-l}{l}\geq 0, \quad \forall\, l\in[K], \tag{82b}$$

which implies

$$R^*_{\mathfrak{A}}(\boldsymbol{m})\geq\max_{\lambda\geq 0}\left\{\min_{l\in[K]}\left\{(K-l)/l+\lambda\big(\text{l-K}m\big)\right\}\right\}, \tag{83}$$

In particular, for $m=t/K$ and $t\in[K]$, we have

$$R^*_{\mathfrak{A}}(\boldsymbol{m})\geq\max_{\lambda\geq 0}\Big\{\min\big\{(K-1)-(t-1)\lambda,\ldots,$$
$$(K-t)/t,\ldots,\lambda K(1-m)\big\}\Big\}=(K-t)/t, \tag{84}$$

since this piecewise linear function is maximized by choosing $\frac{K}{t(t+1)}\leq\lambda^*\leq\frac{K}{t(t-1)}$. In general, for $m=(t+\theta)/K$ and $0\leq\theta\leq 1$, we get

$$R^*_{\mathfrak{A}}(\boldsymbol{m})\geq\max_{\lambda\geq 0}\Big\{\min\Big\{\ldots,\frac{K-t}{t}-\theta\lambda,$$
$$\frac{K-t-1}{t+1}-(1-\theta)\lambda,\ldots\Big\}\Big\}, \tag{85}$$

$$=\frac{K-t}{t}-\frac{\theta K}{t(t+1)}=\frac{K-t}{t}-\frac{(Km-t)K}{t(t+1)}, \tag{86}$$

which is equal to (13).

C. Converse Proof of Theorem 4

Similarly, for $t\leq\sum_{j=1}^{K}m_j\leq t+1$ and $\alpha_{\boldsymbol{q}}=1/(K-1)!$, the lower bound simplifies to

$$R^*_{\mathfrak{A}}(\boldsymbol{m})\geq\max_{\lambda_0\in\mathbb{R},\lambda_j\geq 0} \ -\lambda_0-\sum_{j=1}^{K}\lambda_j m_j \tag{87a}$$

$$\text{subject to } \lambda_0+\sum_{i\in\mathcal{S}}\lambda_i+\frac{K-l}{l}\geq 0, \quad \forall\, l\in[K], \tag{87b}$$

In turn, by choosing $\lambda_j=\lambda$, $\forall j$, we get

$$R^*_{\mathfrak{A}}(\boldsymbol{m})\geq\max_{\lambda\geq 0}\Big\{\min_{l\in[K]}\Big\{(K-l)/l+\lambda\Big(l-\sum_{j=1}^{K}m_j\Big)\Big\}\Big\}. \tag{88}$$

In particular, for $\sum_{j=1}^{K}m_j=(t+\theta)$ and $0\leq\theta\leq 1$, we get

$$R^*_{\mathfrak{A}}(\boldsymbol{m})\geq\max_{\lambda\geq 0}\Big\{\min\Big\{\ldots,\frac{K-t}{t}-\theta\lambda,$$
$$\frac{K-t-1}{t+1}-(1-\theta)\lambda,\ldots\Big\}\Big\}, \tag{89}$$

$$=\frac{K-t}{t}-\frac{\theta K}{t(t+1)}, \tag{90}$$

$$=\frac{tK+(t+1)(K-t)}{t(t+1)}-\frac{K\sum_{j=1}^{K}m_j}{t(t+1)}. \tag{91}$$

D. Converse Proof of Theorem 5

By substituting, $\alpha_{\boldsymbol{q}}=1$ for $j\in[l]$, $\boldsymbol{q}=[1,2,\ldots,j-1,j+1,\ldots,K]$, and $\alpha_{\boldsymbol{q}}=1/(K-l-1)!$ for $j\in\{l+1,\ldots,K\}$, $\boldsymbol{q}=[1,\ldots,l,\boldsymbol{x}]$, $\forall\boldsymbol{x}\in\mathcal{P}_{\{l+1,\ldots,K\}\setminus\{j\}}$, in Theorem 2, we get

$$\gamma_{\mathcal{S}}\triangleq\begin{cases}K-1, \text{ for } |\mathcal{S}|=1,\\ \dfrac{K+l(|\mathcal{S}|-1)+|\mathcal{S}|}{|\mathcal{S}|}, \quad \text{for } \mathcal{S}\subset\{l+1,\ldots,K\}\\ \qquad\qquad\qquad\qquad \text{and } 2\leq|\mathcal{S}|\leq K-1,\\ \min\limits_{i\in\mathcal{S}} i-1, \quad \text{for } \mathcal{S}\cap[l]\neq\phi \text{ and } 2\leq|\mathcal{S}|\leq K-1,\\ 0, \quad \text{for } \mathcal{S}=[K].\end{cases} \tag{92}$$

In particular, for $\mathcal{S}\subset\{l+1,\ldots,K\}$ and $2\leq|\mathcal{S}|\leq K-1$, we have

$$\gamma_{\mathcal{S}}=\sum_{i=l}^{K-|\mathcal{S}|}\frac{i(i-l)!\,(|\mathcal{S}|-1)}{(K-l-1)!}\binom{K-l-|\mathcal{S}|}{i-l}(K-i-2)!, \tag{93}$$

$$=\frac{K+l(|\mathcal{S}|-1)+|\mathcal{S}|}{|\mathcal{S}|}, \tag{94}$$

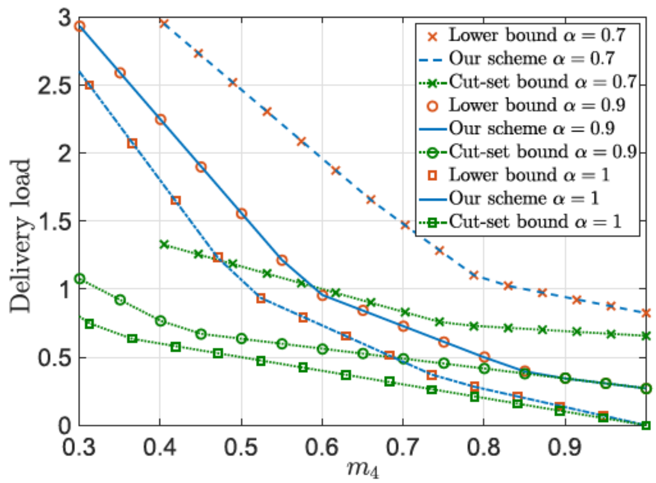


Fig. 5. Comparing $R^*_{\mathfrak{A},\mathfrak{D}}(\boldsymbol{m})$, lower bound on $R^*_{\mathfrak{A}}(\boldsymbol{m})$, and cut-set bound in (95), for $K = N = 4$, and $m_k = \alpha\, m_{k+1}$.

where (93) follows from the number of vectors $\boldsymbol{q} \in \mathcal{P}_{[K]\setminus\{j\}}$ such that $q_k = k, \forall k \in [l]$, $q_{l+1} \in \mathcal{S}$, and $\{q_{l+1}, \ldots, q_i\} \cap \mathcal{S} = \phi$. More specifically, there are $\binom{K-l-|\mathcal{S}|}{i-l}(i-l)!$ choices for $\{q_{l+1}, \ldots, q_i\}$, $(|\mathcal{S}|-1)$ choices for q_{i+1}, and $(K-i-2)!$ choices for elements in $[K] \setminus (\{j\} \cup \{q_1, \ldots, q_{i+1}\})$.

In turn, based on (92), we can verify that $\lambda_0 = -(3K - l - 2)/2$, $\lambda_j = K - j$ for $j \in [l]$, and $\lambda_j = (K-l)/2$ for $j \in \{l+1, \ldots, K\}$, is a feasible solution to (11).

Remark 4: In this region, we achieve the tightest lower bound by choosing $\alpha_{\boldsymbol{q}}$, taking into consideration that the delivery load depends on the individual cache sizes of the users in $[l]$ and the aggregate cache size of the users in $\{l+1, \ldots, K\}$.

VII. Discussion

A. The D2D Delivery Load Memory Trade-Off

In Section III-B, we have characterized the D2D delivery load memory trade-off with uncoded placement and one-shot delivery, $R^*_{\mathfrak{A}}(\boldsymbol{m})$, for several special cases.

For general systems, we observe numerically that the proposed caching scheme coincides with the lower bound in Theorem 2. For example, in Fig. 5, we compare the D2D delivery load $R^*_{\mathfrak{A},\mathfrak{D}}(\boldsymbol{m})$ achievable with our proposed caching scheme with the lower bound on $R^*_{\mathfrak{A}}(\boldsymbol{m})$ in Theorem 2, for $K = N = 4$ and $m_k = \alpha\, m_{k+1}$, and observe they coincide. We also compare the achievable delivery load with a straight forward generalization of the cut-set bound in [6] for unequal caches, which given by

$$R^*(\boldsymbol{m}, N) \geq \max_{s \in [K]} \left\{ s - N \frac{\sum_{i=1}^{s} m_i}{\lfloor N/s \rfloor} \right\}. \qquad (95)$$

From Fig. 5, we observe that in general a gap exists between the cut-set bound in (95) and $R^*_{\mathfrak{A}}(\boldsymbol{m})$, except for the case in Theorem 6.

B. Comparison Between Server-Based and D2D-Based Delivery Loads

By comparing the server-based system [4], [15] delivery load and D2D-based system delivery load, we observe the following:

- The D2D-based delivery load memory trade-off with uncoded placement and one-shot delivery, $R^*_{\mathfrak{A},\text{D2D}}(K, \frac{m_{\text{tot}}}{K})$, for a system with K users and equal cache size $m = m_{\text{tot}}/K$, is equal to the server-based delivery load memory trade-off assuming uncoded placement for a system with $K-1$ users and cache size $m = (m_{\text{tot}} - 1)/(K-1)$, which we denote by $R^*_{\mathfrak{A},\text{Ser}}(K-1, \frac{m_{\text{tot}}-1}{K-1})$ [4]. In particular, for $m_{\text{tot}} \in [K]$, we have

$$R^*_{\mathfrak{A},\text{Ser}}\Big(K-1, \frac{m_{\text{tot}}-1}{K-1}\Big) = \frac{(K-1)(1-\frac{m_{\text{tot}}-1}{K-1})}{1+(K-1)(\frac{m_{\text{tot}}-1}{K-1})} = \frac{1-\frac{m_{\text{tot}}}{K}}{\frac{m_{\text{tot}}}{K}}$$
$$= R^*_{\mathfrak{A},\text{D2D}}\Big(K, \frac{m_{\text{tot}}}{K}\Big). \qquad (96)$$

- From Theorem 4, we conclude that if $m_{\text{tot}} \triangleq \sum_{k=1}^{K} m_k$ and $m_1 \geq (m_{\text{tot}}-1)/(K-1)$, then the D2D delivery load memory trade-off with uncoded placement and one-shot delivery, $R^*_{\mathfrak{A},\text{D2D}}(K, \boldsymbol{m})$, for a system with K users and distinct cache sizes $\boldsymbol{m}$, is equal to $R^*_{\mathfrak{A},\text{D2D}}(K, \frac{m_{\text{tot}}}{K})$. In turn, if $m_1 \geq (m_{\text{tot}}-1)/(K-1)$, then $R^*_{\mathfrak{A},\text{D2D}}(K, \boldsymbol{m}) = R^*_{\mathfrak{A},\text{Ser}}(K-1, \frac{m_{\text{tot}}-1}{K-1})$.
- For a K-user D2D system with $m_K = 1$, user K has access to the whole library and is able to deliver all the missing pieces to the other users. In turn, the D2D delivery load $R^*_{\mathfrak{A},\text{D2D}}(K, [m_1, \ldots, m_{K-1}, 1])$ is equal to $R^*_{\mathfrak{A},\text{Ser}}(K-1, [m_1, \ldots, m_{K-1}])$. For example, for $K=3$, we have

$$R^*_{\mathfrak{A},\text{D2D}}(3, [m_1, m_2, 1]) = R^*_{\mathfrak{A},\text{Ser}}(2, [m_1, m_2])$$
$$= \max\{2-2m_1-m_2,\ 1-m_1\}. \quad (97)$$

C. Non-Uniform File Popularity

Previous works on non-uniform file popularity [33]–[37] have considered minimizing the average delivery load over all possible demands in the shared-bottleneck model [4]. Different strategies for grouping the files according to their popularity have been proposed in [33]–[37]. In particular, reference [35] has shown that dividing the files into two groups and caching only the group of popular files is order-optimal. The scheme in [35] and our proposed scheme can be combined, where we only consider the most popular files. However, the server may need to participate in the delivery phase in order to deliver the file pieces that are not cached by any user. Analyzing this trade-off between the D2D delivery load and the delivery load on the server is an interesting future research direction.

D. Connection Between Coded Distributed Computing and D2D Coded Caching Systems

In coded distributed computing (CDC) systems, the computation of a function over the distributed computing nodes is executed in two stages, named *Map* and *Reduce* [39]. In the former, each computing node maps its local inputs to a set of intermediate values. In order to deliver the intermediate values required for computing the final output at each node, the nodes create multicast transmissions by exploiting the redundancy in computations at the nodes. In the latter, each node reduces the intermediate values retrieved from the multicast signals and the local intermediate values to the desired final outputs.

For CDC systems where the nodes are required to compute different final outputs and each of the final outputs is computed by one node only, the CDC problem can be mapped to a D2D coded caching problem, where the cache placement scheme is uncoded and symmetric over the files [39], [41]. Therefore, the D2D caching scheme proposed in this work can be utilized in heterogeneous CDC systems where the nodes have varying computational/storage capabilities [42]. The mapping between the two problems is described in the following remark.

Remark 5: A D2D caching system with K users, N files, each with size F symbols, where m_k is the normalized cache size at user k, corresponds to a CDC system with K nodes, F files, N final outputs, where $\tilde{M}_k = m_k F$ is the number of files stored at node k. More specifically, in the map stage, node k computes N intermediate values for each cached file. In the reduce stage, node k computes N/K final outputs from the local intermediate values combined with those retrieved from the multicast signals.

Remark 6: Reference [42] derived the optimal communication load in a heterogeneous CDC system consisting of three nodes with different computational/storage capabilities. As a consequence of Remark 5, the optimal communication load found in [42] is the same as the minimum worst-case D2D delivery load with uncoded placement in Theorem 7.

VIII. Conclusion

In this paper, we have proposed a coded caching scheme that minimizes the worst-case delivery load for D2D-based content delivery to users with unequal cache sizes. We have derived a lower bound on the delivery load with uncoded placement and one-shot delivery. We have proved the optimality of our delivery scheme for several cases of interest. In particular, we explicitly characterize $R^*_{\mathfrak{A}}(m)$ for the following cases: (i) $m_k = m, \forall k$, (ii) $(K-2)m_1 \geq \sum_{k=2}^K m_k - 1$, (iii) $\sum_{k=1}^K m_k \leq 2$, (iv) $\sum_{k=1}^K m_k \geq K-1$, and (v) $K=3$. More specifically, for $m_k = m, \forall k$, we have shown the optimality of the caching scheme in [6]. We have also shown that the minimum delivery load depends on the sum of the cache sizes and not the individual cache sizes if the smallest cache size satisfies $(K-2)m_1 \geq \sum_{k=2}^K m_k - 1$.

In the small total memory regime where $\sum_{k=1}^K m_k \leq 2$, we have shown that there exist $K-1$ levels of heterogeneity and in the lth heterogeneity level $R^*_{\mathfrak{A}}(m)$ depends on the individual cache sizes of users $\{1,\dots,l\}$ and the sum of the cache sizes of remaining users. In the large total memory regime where $\sum_{k=1}^K m_k \geq K-1$ and $(K-2)m_1 < \sum_{k=2}^K m_k - 1$, we have shown that our caching scheme achieves the minimum delivery load assuming general placement and delivery. That is, it coincides with the cut-set bound [6]. We have articulated the relationship between the server-based and D2D delivery problems. Finally, we have discussed the coded distributed computing (CDC) problem [39] and how our proposed D2D caching scheme can be tailored for heterogeneous CDC systems where the nodes have unequal storage.

Future directions include considering multi-shot schemes that utilize previous transmitted signals in delivery, heterogeneity in cache sizes and node capabilities for hierarchical cache-enabled networks, and general network topologies.

Appendix A
Achievability Proof of Theorem 7

Region I: $1 \leq m_1 + m_2 + m_3 \leq 2$ *and* $m_1 \geq m_2 + m_3 - 1$

In this region, we show that there exists a feasible solution to (5) that achieves $R^*_{\mathfrak{A},\mathfrak{D}}(m) = \frac{7}{2} - \frac{3}{2}(m_1 + m_2 + m_3)$. In particular, we consider the caching schemes described by $v_{1\to\{2\}} = v_{1\to\{3\}} = a_{\{1\}}$, $v_{2\to\{1\}} = v_{2\to\{3\}} = a_{\{2\}}$, $v_{3\to\{1\}} = v_{3\to\{2\}} = a_{\{3\}}$, $v_{1\to\{2,3\}} + v_{2\to\{1,3\}} = a_{\{1,2\}}$, $v_{1\to\{2,3\}} + v_{3\to\{1,2\}} = a_{\{1,3\}}$, $v_{2\to\{1,3\}} + v_{3\to\{1,2\}} = a_{\{2,3\}}$, and $a_{\{1,2,3\}} = 0$. In turn, the placement feasibility conditions in (18) reduce to

$$v_{1\to\{2,3\}} + v_{2\to\{1,3\}} + v_{3\to\{1,2\}} = \frac{m_1+m_2+m_3-1}{2}, \quad (98a)$$

$$a_{\{1\}} + v_{1\to\{2,3\}} = \frac{m_1+1-m_2-m_3}{2}, \quad (98b)$$

$$a_{\{2\}} + v_{2\to\{1,3\}} = \frac{m_2+1-m_1-m_3}{2}, \quad (98c)$$

$$a_{\{3\}} + v_{3\to\{1,2\}} = \frac{m_3+1-m_1-m_2}{2}. \quad (98d)$$

Note that any caching scheme satisfying (98), achieves the D2D delivery load

$$R^*_{\mathfrak{A},\mathfrak{D}}(m) = 2\big(a_{\{1\}} + a_{\{2\}} + a_{\{3\}}\big) + v_{1\to\{2,3\}} + v_{2\to\{1,3\}} + v_{3\to\{1,2\}} = \frac{7}{2} - \frac{3}{2}(m_1+m_2+m_3). \quad (99)$$

In turn, we only need to choose a non-negative solution to (98), for instance we can choose $a_{\{j\}} = \rho_j\big(2 - m_1 - m_2 - m_3\big)$, such that $\sum_{j=1}^3 \rho_j = 1$, and $0 \leq \rho_j \leq \dfrac{2\, m_j + 1 - \sum_{i=1}^3 m_i}{2\big(2 - \sum_{i=1}^3 m_i\big)}$.

Region II: $1 \leq m_1 + m_2 + m_3 \leq 2$ *and* $m_1 < m_2 + m_3 - 1$

In this region, we achieve the D2D delivery load $R^*_{\mathfrak{A},\mathfrak{D}}(m) = 3 - 2\, m_1 - m_2 - m_3$, by considering the caching schemes described by $v_{1\to\{2\}} = v_{1\to\{3\}} = a_{\{1\}} = 0$, $v_{2\to\{1\}} = v_{2\to\{3\}} = a_{\{2\}}$, $v_{3\to\{2\}} = a_{\{3\}}$, $v_{3\to\{1\}} = a_{\{3\}} + \big(a_{\{2,3\}} - a_{\{1,2\}} - a_{\{1,3\}}\big)$, $v_{1\to\{2,3\}} = 0$, $v_{2\to\{1,3\}} = a_{\{1,2\}}$, $v_{3\to\{1,2\}} = a_{\{1,3\}}$, $a_{\{2,3\}} = m_2 + m_3 - 1$ and $a_{\{1,2,3\}} = 0$. Hence, we only need to choose a non-negative solution to the following equations

$$a_{\{2\}} + a_{\{1,2\}} = 1 - m_3, \quad (100)$$

$$a_{\{3\}} + a_{\{1,3\}} = 1 - m_2, \quad (101)$$

$$a_{\{1,2\}} + a_{\{1,3\}} = m_1, \quad (102)$$

which follows from (18). Note that any non-negative solution to (100), achieves $R^*_{\mathfrak{A},\mathfrak{D}}(m) = 3 - 2\, m_1 - m_2 - m_3$. For instance, we can choose $a_{\{1,3\}} = 0$ when $m_1 + m_3 \leq 1$ and $a_{\{2\}} = 0$ when $m_1 + m_3 > 1$.

Region III: $m_1 + m_2 + m_3 > 2$ *and* $m_2 + m_3 \leq 1 + m_1$

In order to achieve $R^*_{\mathfrak{A},\mathfrak{D}}(m) = \frac{3}{2} - \frac{1}{2}\big(m_1 + m_2 + m_3\big)$, we consider the caching scheme described by $v_{1\to\{2,3\}} = (m_1 + 1 - m_2 - m_3)/2$, $v_{2\to\{1,3\}} = (m_2 + 1 - m_1 - m_3)/2$, $v_{3\to\{1,2\}} = (m_3 + 1 - m_1 - m_2)/2$, $a_{\{1,2\}} = 1 - m_3$, $a_{\{1,3\}} = 1 - m_2$, $a_{\{2,3\}} = 1 - m_1$, and $a_{\{1,2,3\}} = m_1 + m_2 + m_3 - 2$.

Region IV: $m_1 + m_2 + m_3 > 2$ *and* $m_2 + m_3 > 1 + m_1$

Finally, $R^*_{\mathfrak{A},\mathfrak{D}}(m) = 1 - m_1$ is achieved by $a_{\{1,2\}} = 1 - m_3$, $a_{\{1,3\}} = 1 - m_2$, $a_{\{2,3\}} = 1 - m_1$, $a_{\{1,2,3\}} = m_1 + m_2 + m_3 - 2$, $v_{3\to\{1\}} = m_2 + m_3 - m_1 - 1$, $v_{2\to\{1,3\}} = 1 - m_3$, and $v_{3\to\{1,2\}} = 1 - m_2$.

Appendix B
Converse Proof of Theorem 7

By substituting $\alpha_q = 1/2, \forall q \in \mathcal{P}_{[3]\setminus\{j\}}, \forall j \in [3]$ in Theorem 2, we get

$$R^*_{\mathfrak{A}}(m) \geq \max_{\lambda_0 \in \mathbb{R}, \lambda_k \geq 0} \; -\lambda_0 - \lambda_1 m_1 - \lambda_2 m_2 - \lambda_3 m_3 \quad (103a)$$

$$\text{subject to } \lambda_0 + \lambda_j + 2 \geq 0, \quad \forall j \in [3], \quad (103b)$$

$$\lambda_0 + \lambda_i + \lambda_j + 1/2 \geq 0, \quad \forall i \in [3], \; j \neq i, \quad (103c)$$

$$\lambda_0 + \lambda_1 + \lambda_2 + \lambda_3 \geq 0. \quad (103d)$$

By choosing two feasible solutions to (103), we get

$$R^*_{\mathfrak{A}}(m) \geq \max\left\{\frac{7}{2} - \frac{3}{2}\big(m_1 + m_2 + m_3\big), \frac{3}{2} - \frac{1}{2}\big(m_1 + m_2 + m_3\big)\right\}. \quad (104)$$

Similarly, by substituting $\alpha_{[2,3]} = \alpha_{[1,3]} = \alpha_{[1,2]} = 1$ in Theorem 2, we can show that

$$R^*_{\mathfrak{A}}(m) \geq \max\{3 - 2m_1 - m_2 - m_3, \; 1 - m_1\}. \quad (105)$$

References

[1] Cisco, "Cisco VNI forecast methodology," Cisco, San Jose, CA, USA, Jun. 2016, pp. 2015–2020.

[2] A. Asadi, Q. Wang, and V. Mancuso, "A survey on device-to-device communication in cellular networks," *IEEE Commun. Surveys Tuts.*, vol. 16, no. 4, pp. 1801–1819, Apr. 2014.

[3] M. A. Maddah-Ali and U. Niesen, "Coding for caching: Fundamental limits and practical challenges," *IEEE Commun. Mag.*, vol. 54, no. 8, pp. 23–29, Aug. 2016.

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

[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] M. Ji, G. Caire, and A. F. Molisch, "Fundamental limits of caching in wireless D2D networks," *IEEE Trans. Inf. Theory*, vol. 62, no. 2, pp. 849–869, Feb. 2016.

[7] Q. Yu, M. A. Maddah-Ali, and A. S. Avestimehr, "The exact rate-memory tradeoff for caching with uncoded prefetching," *IEEE Trans. Inf. Theory*, vol. 64, no. 2, pp. 1281–1296, Feb. 2018.

[8] K. Wan, D. Tuninetti, and P. Piantanida, "On the optimality of uncoded cache placement," in *Proc. IEEE Inf. Theory Workshop (ITW)*, Sep. 2016, pp. 161–165.

[9] K. Wan, D. Tuninetti, and P. Piantanida, "A novel index coding scheme and its application to coded caching," Feb. 2017, *arXiv:1702.07265*. [Online]. Available: https://arxiv.org/abs/1702.07265

[10] S. Wang, W. Li, X. Tian, and H. Liu, "Coded caching with heterogenous cache sizes," Apr. 2015, *arXiv:1504.01123*. [Online]. Available: https://arxiv.org/abs/1504.01123

[11] M. Mohammadi Amiri, Q. Yang, and D. Gunduz, "Decentralized caching and coded delivery with distinct cache capacities," *IEEE Trans. Commun.*, vol. 65, no. 11, pp. 4657–4669, Nov. 2017.

[12] A. Sengupta, R. Tandon, and T. C. Clanc, "Layered caching for heterogeneous storage," in *Proc. 50th Asilomar Conf. Signals, Syst. Comput.*, Nov. 2016, pp. 719–723.

[13] A. M. Ibrahim, A. A. Zewail, and A. Yener, "Centralized coded caching with heterogeneous cache sizes," in *Proc. IEEE Wireless Commun. Netw. Conf. (WCNC)*, Mar. 2017, pp. 1–6.

[14] A. M. Ibrahim, A. A. Zewail, and A. Yener, "Optimization of heterogeneous caching systems with rate limited links," in *Proc. IEEE Int. Conf. Commun.(ICC)*, May 2017, pp. 1–6.

[15] A. M. Ibrahim, A. A. Zewail, and A. Yener, "Coded caching for heterogeneous systems: An optimization prespective," *IEEE Trans. Commun.*, vol. 67, no. 8, pp. 5321–5335, Aug. 2019.

[16] F. Arbabjolfaei, B. Bandemer, Y.-H. Kim, E. Sasoglu, and L. Wang, "On the capacity region for index coding," in *Proc. IEEE Int. Symp. Inf. Theory*, Jul. 2013, pp. 962–966.

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

[18] M. Ji, R.-R. Chen, G. Caire, and A. F. Molisch, "Fundamental limits of distributed caching in multihop D2D wireless networks," in *Proc. IEEE Int. Symp. Inf. Theory (ISIT)*, Jun. 2017, pp. 2950–2954.

[19] A. Shabani, S. P. Shariatpanahi, V. Shah-Mansouri, and A. Khonsari, "Mobility increases throughput of wireless device-to-device networks with coded caching," in *Proc. IEEE Int. Conf. Commun. (ICC)*, May 2016, pp. 1–6.

[20] A. Tebbi and C. W. Sung, "Coded caching in partially cooperative D2D communication networks," Sep. 2017, *arXiv:1709.06281*. [Online]. Available: https://arxiv.org/abs/1709.06281

[21] A. K. Chorppath, J. Hackel, and F. H. Fitzek, "Network coded caching and D2D cooperation in wireless networks," in *Proc. VDE EW*, May 2017, pp. 1–6.

[22] Z. H. Awan and A. Sezgin, "Fundamental limits of caching in D2D networks with secure delivery," in *Proc. IEEE Int. Conf. Commun. Workshop (ICCW)*, Jun. 2015, pp. 464–469.

[23] A. A. Zewail and A. Yener, "Device-to-device secure coded caching," *IEEE Trans. Inf. Forensics Security*, vol. 15, pp. 1513–1524, Sep. 2019.

[24] C. C. Yapar, K. Wan, R. F. Schaefer, and G. Caire, "On the optimality of D2D coded caching with uncoded cache placement and one-shot delivery," Jan. 2019, *arXiv:1901.05921*. [Online]. Available: https://arxiv.org/abs/1901.05921

[25] P. Gupta and P. R. Kumar, "The capacity of wireless networks," *IEEE Trans. Inf. Theory*, vol. 46, no. 2, pp. 388–404, Mar. 2000.

[26] Q. Yang and D. Gunduz, "Coded caching and content delivery with heterogeneous distortion requirements," *IEEE Trans. Inf. Theory*, vol. 64, no. 6, pp. 4347–4364, Jun. 2018.

[27] A. M. Daniel and W. Yu, "Optimization of heterogeneous coded caching," Aug. 2017, *arXiv:1708.04322*. [Online]. Available: https://arxiv.org/abs/1708.04322

[28] D. Cao, D. Zhang, P. Chen, N. Liu, W. Kang, and D. Gündüz, "Coded caching with heterogeneous cache sizes and link qualities: The two-user case," Feb. 2018, *arXiv:1802.02706*. [Online]. Available: https://arxiv.org/abs/1802.02706

[29] J. Zhang, X. Lin, C.-C. Wang, and X. Wang, "Coded caching for files with distinct file sizes," in *Proc. IEEE Int. Symp. Inf. Theory (ISIT)*, Jun. 2015, pp. 1686–1690.

[30] 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.

[31] P. Hassanzadeh, E. Erkip, J. Llorca, and A. Tulino, "Distortion-memory tradeoffs in cache-aided wireless video delivery," in *Proc. 53rd Annu. Allerton Conf. Commun., Control, Comput. (Allerton)*, Sep. 2015, pp. 1150–1157.

[32] A. M. Ibrahim, A. A. Zewail, and A. Yener, "On coded caching with heterogeneous distortion requirements," in *Proc. Inf. Theory Appl. Workshop (ITA)*, Feb. 2018, pp. 1–9.

[33] U. Niesen and M. A. Maddah-Ali, "Coded caching with nonuniform demands," in *Proc. IEEE Conf. Comput. Commun. Workshops (INFOCOM WKSHPS)*, Apr./May 2014, pp. 221–226.

[34] 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.

[35] 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.

[36] J. Hachem, N. Karamchandani, and S. Diggavi, "Effect of number of users in multi-level coded caching," in *Proc. IEEE Int. Symp. Inf. Theory (ISIT)*, Jun. 2015, pp. 1701–1705.

[37] J. Hachem, N. Karamchandani, and S. Diggavi, "Coded caching for multi-level popularity and access," *IEEE Trans. Inf. Theory*, to be published.

[38] S. Jin, Y. Cui, H. Liu, and G. Caire, "Structural properties of uncoded placement optimization for coded delivery," Jul. 2017, *arXiv:1707.07146*. [Online]. Available: https://arxiv.org/abs/1707.07146

[39] S. Li, M. A. Maddah-Ali, Q. Yu, and A. S. Avestimehr, "A fundamental tradeoff between computation and communication in distributed computing," *IEEE Trans. Inf. Theory*, vol. 64, no. 1, pp. 109–128, Jan. 2018.

[40] Y.-J. Ku *et al.*, "5G radio access network design with the fog paradigm: Confluence of communications and computing," *IEEE Commun. Mag.*, vol. 55, no. 4, pp. 46–52, Apr. 2017.

[41] K. Wan, D. Tuninetti, M. Ji, G. Caire, and P. Piantanida, "Fundamental limits of decentralized data shuffling," Jun. 2018, *arXiv:1807.00056*. [Online]. Available: https://arxiv.org/abs/1807.00056

[42] M. Kiamari, C. Wang, and A. S. Avestimehr, "On heterogeneous coded distributed computing," Sep. 2017, *arXiv:1709.00196*. [Online]. Available: https://arxiv.org/abs/1709.00196

Abdelrahman M. Ibrahim (Member, IEEE) received the B.Sc. degree in electrical engineering from Alexandria University, Alexandria, Egypt, in 2011, the M.Sc. degree in wireless communications from Nile University, Giza, Egypt, in 2014, and the Ph.D. degree in electrical engineering from Pennsylvania State University, University Park, PA, USA, in 2019. His Ph.D. dissertation is focused on data storage and energy management in emerging networks. He was an Exchange Research Assistant with Sabanci University, Istanbul, Turkey, from 2013 to 2014, funded by the Marie Curie International Research Exchange Scheme. From 2014 to 2019, he was a Research Assistant with the Wireless Communications and Networking (WCAN) Laboratory, Pennsylvania State University. He is currently a Senior Systems Engineer with the Department of Wireless R&D, Qualcomm Technologies Inc., San Diego, CA, USA. His research interests include 5G New Radio (NR) PHY and MAC layers, data storage systems, cache-aided networks, green communications, and resource allocation in wireless networks.

Ahmed A. Zewail (Member, IEEE) received the B.Sc. degree (Hons.) in electrical engineering from Alexandria University, Alexandria, Egypt, in 2011, the M.Sc. degree in wireless communications from Nile University, Giza, Egypt, in 2013, and the Ph.D. degree from Pennsylvania State University, in 2019. His M.Sc. thesis focused on the capacity and degrees of freedom of relay networks. His Ph.D. dissertation focused on secrecy guarantees in emerging networks, e.g., untrusted relay networks and cache-aided networks. From 2011 to 2013, he was a Research Assistant with the Wireless Intelligence Networks Center (WINC), Nile University. From 2013 to 2019, he was a Research Assistant with the Wireless Communications and Networking (WCAN) Laboratory, Pennsylvania State University. He is currently with Qualcomm Technologies Inc., San Diego, CA, USA. His research interests include network information theory, cache-aided networks, wireless communications, and physical-layer security. His graduation project was about developing warehouse management systems using RFID and received the first place in INDAC-Siemens 2011 Competition.

Aylin Yener (Fellow, IEEE) received the B.Sc. degree in electrical and electronics engineering and the B.Sc. degree in physics from Bogazici University, Istanbul, Turkey, and the M.S. and Ph.D. degrees in electrical and computer engineering from the Wireless Information Network Laboratory (WINLAB), Rutgers University, New Brunswick, NJ, USA. She holds the Roy and Lois Chope Chair in Engineering at The Ohio State University, Columbus Ohio, since 2020, where she is a Professor of electrical and computer engineering, Professor of integrated systems engineering, and Professor of computer science and engineering. Until 2020, she was a University Distinguished Professor of electrical engineering and a Dean's Fellow at The Pennsylvania State University, University Park, PA, USA, where she joined the faculty as an Assistant Professor in 2002. She was a Visiting Professor of Electrical Engineering at Stanford University in 2016–2018 and a Visiting Associate Professor in the same department in 2008–2009. Her current research interests are in information security, green communications, caching systems, and more generally in the fields of information theory, communication theory and networked systems. She received the NSF CAREER Award in 2003, the Best Paper Award in Communication Theory from the IEEE International Conference on Communications in 2010, the Penn State Engineering Alumni Society (PSEAS) Outstanding Research Award in 2010, the IEEE Marconi Prize Paper Award in 2014, the PSEAS Premier Research Award in 2014, the Leonard A. Doggett Award for Outstanding Writing in Electrical Engineering at Penn State in 2014, the IEEE Women in Communications Engineering Outstanding Achievement Award in 2018, and the IEEE Communications Society Best Tutorial Paper Award in 2019. She has been a Distinguished Lecturer for the IEEE Information Theory Society (2019–2020), the IEEE Communications Society (2018–2019) and the IEEE Vehicular Technology Society (2017–2021).

Dr. Yener is serving as the President of the IEEE Information Theory Society in 2020. Previously, she was the Vice President (2019), the Second Vice President (2018), an elected member of the Board of Governors (2015–2018), and the Treasurer (2012–2014) of the IEEE Information Theory Society. She served as the Student Committee Chair for the IEEE Information Theory Society (2007–2011), and was the Co-Founder of the Annual School of Information Theory in North America in 2008. She was a Technical (Co)-Chair for various symposia/tracks at the IEEE ICC, PIMRC, VTC, WCNC, and Asilomar in 2005, 2008–2014 and 2018. Previously, she served as an Editor for IEEE TRANSACTIONS ON COMMUNICATIONS (2009–2012), an Editor for IEEE TRANSACTIONS ON MOBILE COMPUTING (2017–2018), and an Editor and an Editorial Advisory Board Member for IEEE TRANSACTIONS ON WIRELESS COMMUNICATIONS (2001–2012). She also served a Guest Editor for IEEE TRANSACTIONS ON INFORMATION FORENSICS AND SECURITY in 2011, and IEEE JOURNAL ON SELECTED AREAS IN COMMUNICATIONS in 2015. Currently, she serves as a Senior Editor for IEEE JOURNAL ON SELECTED AREAS IN COMMUNICATIONS and is on the Senior Editorial Board of IEEE JOURNAL ON SELECTED AREAS IN INFORMATION THEORY.