Device-to-Device Coded Caching with Heterogeneous Cache Sizes

Abdelrahman M. Ibrahim, Ahmed A. Zewail, and Aylin Yener

Wireless Communications and Networking Laboratory (WCAN)
School of Electrical Engineering and Computer Science
The Pennsylvania State University, University Park, PA 16802.
{*ami137,zewail*}*@psu.edu* *yener@engr.psu.edu*

***Abstract*—This paper considers a device-to-device (D2D) coded caching system where the users have differing cache sizes. During low traffic hours, the server places subsets of the files at the users' cache memories, in a manner that enables serving the users' requests via D2D transmissions during peak traffic hours. The objective is to jointly design the users' cache contents and the D2D transmissions in order to minimize the D2D delivery load. In particular, we seek to identify the optimal uncoded placement and linear delivery schemes. We propose a novel lower bound on the D2D delivery load under uncoded placement, which enables us to explicitly characterize the minimum D2D delivery load under uncoded placement for several cases of interest.**

I. Introduction

In order to cope with the ever-increasing wireless data traffic, e.g., video-on-demand services [1], it is imperative to develop novel techniques that fully utilize network resources. Caching [2]–[10] and device-to-device (D2D) communications [11] are proposed to alleviate network congestion by utilizing the users' cache memories and the radio interface enabling the nodes to directly communicate with each other. In particular, coded caching not only shifts some of the network traffic to off-peak hours, but also creates multicast opportunities that reduce the delivery load on the server [2].

References [3]–[5] have studied the fundamental limits of coded caching in a device-to-device setup. In particular, reference [3] has proposed centralized and decentralized caching schemes for homogeneous systems where the users have equal cache sizes. Reference [4] has studied secure D2D delivery in the presence of an eavesdropper that overhears the D2D transmissions. Reference [5] has considered secure D2D coded caching under confidentiality constraints when users cannot recover the files requested by other users.

Uncoded placement is a class of cache placement schemes adopted in [2], [3], [6]–[10], in which the server partitions the files into pieces and places these pieces at the users' cache memories. The optimality of the caching scheme in [2] under uncoded placement has been studied in [6]–[8]. In particular, references [6], [7] have illustrated that the server-based delivery problem in [2] is equivalent to an index-coding problem and the delivery load in [2] is lower bounded by the acyclic index-coding bound [12, Corollary 1].

In order to account for the distinct storage capabilities of the users in cache-enabled networks, recent reference [9] has studied the server-based delivery problem in systems where users have different cache sizes. In particular, 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 in a centralized caching network, where the delivery phase is also facilitated by the server. Different from [9], in this paper, we consider a D2D delivery model, in which the server is silent, i.e., the users are served via D2D transmissions in the delivery phase, and investigate the impact of heterogeneous cache sizes at the end users. We jointly design cache placement and D2D delivery schemes that minimize the D2D delivery load. In particular, a linear program determines the partitioning of the files in the placement phase and the size and structure of the D2D transmissions.

Additionally, building on [6]–[8], we propose a lower bound on the worst-case D2D delivery load under uncoded placement, which is also defined by a linear program. Using the proposed lower bound, we prove the optimality of the caching scheme in [3] under uncoded placement and characterize explicitly the delivery load memory trade-off under uncoded placement for several cases. In particular, we show that the D2D delivery load depends only on the total cache sizes in the system whenever the smallest cache size is greater than a certain threshold. Additionally, we characterize the trade-off in the case where the total memory is less than double the library size and the case where the total memory is greater than $K-1$ times the library size, where K is the number of users. In turn, we completely characterize the trade-off for the three-user case. Finally, we show numerically that the proposed delivery scheme achieves the minimum D2D delivery load.

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., $\mathcal{P}_{\{1,2\}} = \{[1,2],\ [2,1]\}$.

II. System Model

We consider a network consisting of one server connected to K users. The server is connected to the users via a shared error-free link, and the users are connected to each other via error-free device-to-device (D2D) communication links,

978-1-5386-3180-5/18/$31.00

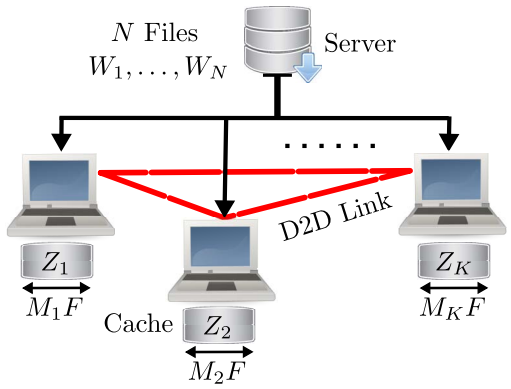

Fig. 1: Caching system with heterogeneous cache sizes.

as illustrated in Fig. 1. The server has access to a library $\{W_1, \ldots, W_N\}$ of N files, each with size F bits. Users are equipped with cache memories that may have different sizes; the size of the cache memory at user k is equal to M_kF bits. Without loss of generality, we let $M_1 \leq M_2 \leq \cdots \leq M_K$. We also 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 case where the number of users is less than the number of files, i.e., $K \leq N$.

Cache-enabled networks have two operational phases [2]. The first phase is called the placement phase, in which the server places subsets of the library at the users' cache memories without the knowledge of the users' demands. In particular, the server places a subset of the library, Z_k, at the cache memory of user k, such that $|Z_k| \leq M_kF$ bits. The second phase is called the delivery phase, in which the users announce their requested files which are represented by the demand vector $\boldsymbol{d} = [d_1, \ldots, d_K]$, i.e., user k requests the file W_{d_k}. The demand vector consists of identical and independent uniform random variables over the files $[N]$ [2].

In the delivery phase, we consider that the users' demands are satisfied using D2D communications only [3], i.e., the server is silent during the delivery phase. This necessitates that the users' cache sizes total at least the library size, i.e., $\sum_{k=1}^{K} m_k \geq 1$. In order to satisfy the demand vector $\boldsymbol{d}$, user j transmits a 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. The transmitted signals $X_{j\to\mathcal{T},\boldsymbol{d}}, j \in [K], \mathcal{T} \subsetneq_\phi [K]\setminus\{j\}$ need to ensure that the requested files are decoded reliably at the users. The delivery load $R_j(\boldsymbol{m})$ refers to the amount of data transmitted by user j as a fraction of the file size F.

Definition 1. *For a given normalized cache size vector $\boldsymbol{m}$, 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 a caching scheme such that*

$$\max_{\boldsymbol{d}\in[N]^K,\, k\in[K]} Pr(\hat{W}_{d_k} \neq W_{d_k}) \leq \epsilon. \tag{1}$$

Furthermore, the infimum over all achievable delivery loads is denoted by $R^*(\boldsymbol{m})$. ■

Note that under uniform demands, user k stores m_kF bits of each file, i.e., equal number of bits are stored per file. Similar to [2], [3], [9], we assume the placement phase to be uncoded and the delivery phase uses linear codes. We denote this class of cache placement and delivery policies by $\mathfrak{A}$ and $\mathfrak{D}$, respectively.

Definition 2. *Under an uncoded placement scheme in $\mathfrak{A}$, and any linear delivery policy in $\mathfrak{D}$, the worst-case D2D delivery load is defined as*

$$R_{\mathfrak{A},\mathfrak{D}} \triangleq \max_{\boldsymbol{d}\in[N]^K} \sum_{j=1}^{K} R_{j,\boldsymbol{d},\mathfrak{A},\mathfrak{D}} = \sum_{j=1}^{K} \sum_{\mathcal{T}\subsetneq_\phi[K]\setminus\{j\}} v_{j\to\mathcal{T}}. \tag{2}$$

By taking the infimum over $\mathfrak{A}$ and all possible delivery policies, we get $R^*_{\mathfrak{A}}(\boldsymbol{m})$. ■

III. Formulation

In this section, we summarize our results on the D2D delivery load memory trade-off. First, in Theorem 1, we show that the minimum worst-case D2D delivery load under uncoded placement and linear delivery, $R^*_{\mathfrak{A},\mathfrak{D}}(\boldsymbol{m})$, can be obtained by solving a linear program. Next, in Theorem 2, we present a lower bound on the D2D delivery load memory trade-off under uncoded placement $R^*_{\mathfrak{A}}(\boldsymbol{m})$, which enable us to characterize the trade-off explicitly for several cases. In particular, for these cases, we show the optimality of linear delivery under uncoded placement, i.e., $R^*_{\mathfrak{A}}(\boldsymbol{m}) = R^*_{\mathfrak{A},\mathfrak{D}}(\boldsymbol{m})$.

Theorem 1. *Given $N \geq K$, and $\boldsymbol{m}$, the achievable minimum worst-case D2D delivery load under uncoded placement and linear delivery, $R^*_{\mathfrak{A},\mathfrak{D}}(\boldsymbol{m})$, is characterized by*

$$\textbf{\textit{O1:}} \quad \min_{\boldsymbol{a},\boldsymbol{u},\boldsymbol{v}} \quad \sum_{j=1}^{K} \sum_{\mathcal{T}\subsetneq_\phi[K]\setminus\{j\}} v_{j\to\mathcal{T}} \tag{3a}$$

$$\textit{subject to} \quad \boldsymbol{a} \in \mathfrak{A}(\boldsymbol{m}), \tag{3b}$$

$$(\boldsymbol{u},\boldsymbol{v}) \in \mathfrak{D}(\boldsymbol{a}), \tag{3c}$$

where $\mathfrak{A}(\boldsymbol{m})$ is the set of uncoded placement schemes defined in (14) and $\mathfrak{D}(\boldsymbol{a})$ is the set of feasible D2D linear delivery schemes defined by (35)-(39). ■

Proof. See Section V. □

The following theorem defines a lower bound on $R^*_{\mathfrak{A}}(\boldsymbol{m})$, which is motivated by the lower bounds on server-based delivery in [6]–[8].

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

$$\textbf{\textit{O2:}} \quad \max_{\lambda_0\in\mathbb{R},\lambda_k\geq 0,\alpha_{\boldsymbol{q}}\geq 0} \quad -\lambda_0 - \sum_{k=1}^{K} m_k\lambda_k \tag{4a}$$

$$\text{subject to} \quad \lambda_0 + \sum_{k\in\mathcal{S}} \lambda_k + \gamma_{\mathcal{S}} \geq 0, \forall\, \mathcal{S} \subsetneq_{\phi} [K], \quad (4b)$$

$$\sum_{\boldsymbol{q}\in\mathcal{P}_{[K]\setminus\{j\}}} \alpha_{\boldsymbol{q}} = 1,\ \forall\, j \in [K], \quad (4c)$$

where $\mathcal{P}_{[K]\setminus\{j\}}$ is the set of all permutations of $[K]\setminus\{j\}$, and

$$\gamma_{\mathcal{S}} \triangleq \begin{cases} K-1, \text{ for } |\mathcal{S}| = 1, \\ \min\limits_{j\in\mathcal{S}} \left\{ \sum\limits_{i=1}^{K-|\mathcal{S}|} \sum\limits_{\substack{\boldsymbol{q}\in\mathcal{P}_{[K]\setminus\{j\}}: \\ q_{i+1}\in\mathcal{S},\{q_1,\ldots,q_i\}\cap\mathcal{S}=\phi}} i\,\alpha_{\boldsymbol{q}} \right\}, \text{ for } 2\leq|\mathcal{S}|\leq K-1 \\ 0, \text{ for } \mathcal{S} = [K]. \end{cases} \quad (5)$$

■

Proof. See Appendix A. □

IV. Results

Next, using Theorem 2, we show the optimality of the caching scheme proposed in [3] under uncoded placement, for systems where the users have equal cache sizes.

Theorem 3. *For $N \geq K$, and $m_k = m, \forall k \in [K]$, the minimum worst-case D2D delivery load under uncoded placement, $R^*_{\mathfrak{A}}(m) = (1-m)/m$, for $m = t/K, t \in [K]$. Furthermore, for $t \leq Km \leq t+1,\ t \in [K-1]$, we have*

$$R^*_{\mathfrak{A}}(m) = \left(\frac{K-t}{t}\right)\big(t+1-Km\big) + \left(\frac{K-t-1}{t+1}\right)\big(Km-t\big). \quad (6)$$

■

Proof. **Achievability**: The centralized caching scheme proposed in [3] achieves (6), which is also the optimal solution of (3). **Converse**: By substituting $\alpha_{\boldsymbol{q}} = 1/(K-1)!$ in Theorem 2, we get $\gamma_{\mathcal{S}} = (K-|\mathcal{S}|)/|\mathcal{S}|$, for $2 \leq |\mathcal{S}| \leq K-1$. In turn,

$$R^*_{\mathfrak{A}}(\boldsymbol{m}) \geq \max_{\lambda\geq 0} \left\{ \min_{l\in[K]} \left\{ \frac{K-l}{l} + \lambda\big(l-Km\big) \right\} \right\} = \frac{K-t}{t}, \quad (7)$$

for $m = t/K$, and $t \in [K]$. Additionally, the delivery load in (6) is achievable for $t \leq Km \leq t+1$, since the objective function in (7) is piecewise linear. □

Corollary 1. *The D2D delivery load memory trade-off under uncoded placement in a K-user system with equal cache sizes m is equal to the server-based delivery load in a $(K-1)$-user system with equal cache sizes $(Km-1)/(K-1)$ [2]. In particular, for $m = t/K, t \in [K]$, we have*

$$R^*_{\mathfrak{A},Ser}\Big(K-1, \frac{Km-1}{K-1}\Big) = \frac{(K-1)(1-\frac{Km-1}{K-1})}{1+(K-1)(\frac{Km-1}{K-1})} = \frac{1-m}{m} = R^*_{\mathfrak{A},D2D}\big(K,m\big). \quad (8)$$

■

The next theorem shows that the heterogeneity in users' cache sizes does not increase the D2D delivery load as long as the smallest cache size m_1 is greater than or equal to $\big(\sum_{k=2}^{K} m_k - 1\big)/(K-2)$. The proof is omitted due to space limitations.

Theorem 4. *For $N \geq K$, $m_1 \leq \ldots \leq m_K$, and $(K-2)m_1 \geq \sum_{k=2}^{K} m_k - 1$, the minimum worst-case D2D delivery load under uncoded placement, $R^*_{\mathfrak{A}}(\boldsymbol{m})$, is given by*

$$\left(\frac{K-t}{t}\right)\left(t+1-\sum_{k=1}^{K} m_k\right) + \left(\frac{K-t-1}{t+1}\right)\left(\sum_{k=1}^{K} m_k - t\right), \quad (9)$$

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

Corollary 2. *Whenever $(K-1)m_1 \geq \sum_{k=1}^{K} m_k - 1$, the D2D delivery load in a K-user system with cache sizes $\boldsymbol{m}$ is equal to the server-based delivery load in a $(K-1)$-user system with equal cache sizes $\big(\sum_{k=1}^{K} m_k - 1\big)/(K-1)$ [2].* ■

The next theorem characterizes the trade-off in the small total memory regime where $\sum\limits_{k=1}^{K} m_k \leq 2$.

Theorem 5. *For $N \geq K$, $m_1 \leq \ldots \leq m_K$, $1 \leq \sum_{k=1}^{K} m_k \leq 2$, $(K-l-1)m_l < \sum\limits_{i=l+1}^{K} m_i - 1$ and $(K-l-2)m_{l+1} \geq \sum\limits_{i=l+2}^{K} m_i - 1$, the minimum worst-case D2D delivery load under uncoded placement, for $l \in [K-2]$ is given as*

$$R^*_{\mathfrak{A}}(\boldsymbol{m}) = \frac{3K-l-2}{2} - \sum_{i=1}^{l}(K-i)m_i - \left(\frac{K-l}{2}\right)\sum_{i=l+1}^{K} m_i. \quad (10)$$

■

Proof. **Achievability**: Proof is omitted due to space limitations. **Converse**: 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, for $2 \leq |\mathcal{S}| \leq K-1$, we get

$$\gamma_{\mathcal{S}} \triangleq \begin{cases} \dfrac{K + l(|S|-1) + |S|}{|S|}, \text{ for } \mathcal{S} \subset \{l+1, \ldots, K\}, \\ \min\limits_{i\in\mathcal{S}} i - 1, \text{ for } \mathcal{S} \cap [l] \neq \phi. \end{cases} \quad (11)$$

In turn, $R^*_{\mathfrak{A}}(\boldsymbol{m})$ is lower bounded by (10). □

In the large total memory regime where $\sum_{k=1}^{K} m_k \geq K-1$, our caching scheme achieves $R^*(\boldsymbol{m})$, which is shown in the next theorem. Achievability proof is omitted due to space limitation. Converse follows from the cut-set bound [3].

Theorem 6. *For $N \geq K$, $m_1 \leq \ldots \leq m_K$, $\sum_{k=1}^{K} m_k \geq K-1$, and $(K-2)m_1 < \sum_{i=2}^{K} m_i - 1$, the minimum worst-case D2D delivery load under uncoded placement,*

$$R^*_{\mathfrak{A}}(\boldsymbol{m}) = R^*(\boldsymbol{m}) = 1 - m_1. \quad (12)$$

■

Finally, combining Theorems 4, 5, and 6, we completely characterize the trade-off under uncoded placement for $K = 3$.

Theorem 7. *For $K = 3$, $N \geq 3$, and $m_1 \leq m_2 \leq m_3$, the minimum worst-case D2D delivery load under uncoded*

placement, $R^*_{\mathfrak{A}}(\boldsymbol{m})$, is given by

$$\max\left\{\frac{7}{2}-\sum_{k=1}^{3}\frac{3m_k}{2}, 3-2m_1-\sum_{k=2}^{3}m_k, \frac{3}{2}-\sum_{k=1}^{3}\frac{m_k}{2}, 1-m_1\right\}. \quad (13)$$

■

Remark 1. *Note that the D2D delivery load in a K-user system with cache sizes $[m_1,\ldots,m_{K-1},1]$ is equal to the server-based delivery load in a $(K-1)$-user system with cache sizes $[m_1,\ldots,m_{K-1}]$ [9]. For example, by substituting $m_3=1$ in (13), we get $R^*_{\mathfrak{A},D2D}(3,[m_1,m_2,1]) = R^*_{\mathfrak{A},Ser}(2,[m_1,m_2]) = \max\{2-2m_1-m_2,\ 1-m_1\}$.* ■

V. Caching Scheme

In this section, we first explain the cache placement phase, in which we require that there is no subfile stored exclusively at the server. Next, we explain the delivery phase which 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 $X_{j\to\mathcal{T}}$[1] to the users in the sets $\mathcal{T}\subsetneq_\phi [K]\setminus\{j\}$.

A. Placement phase

The server partitions each file W_l into 2^K-1 subfiles, $\tilde{W}_{l,\mathcal{S}}, \mathcal{S}\subsetneq_\phi [K]$, such that $\tilde{W}_{l,\mathcal{S}}$ denotes a subset of W_l which is stored exclusively at the users in the set $\mathcal{S}$. The partitioning is symmetric over the files, i.e., $|\tilde{W}_{l,\mathcal{S}}| = a_{\mathcal{S}}F$ bits, $\forall l\in[N]$, where the *allocation variable* $a_{\mathcal{S}}\in[0,1]$ defines the size of $\tilde{W}_{l,\mathcal{S}}$ as a fraction of the file size F. Therefore, the set of feasible uncoded placement schemes, $\mathfrak{A}(\boldsymbol{m})$, is defined by

$$\left\{\boldsymbol{a}\in[0,1]^{2^K}\ \middle|\ \sum_{\mathcal{S}\subsetneq_\phi[K]} a_{\mathcal{S}}=1, \sum_{\mathcal{S}\subset[K]\,:\,k\in\mathcal{S}} a_{\mathcal{S}}\le m_k, \forall k\in[K]\right\}, \quad (14)$$

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

$$Z_k = \bigcup_{l\in[N]}\ \bigcup_{\mathcal{S}\subset[K]\,:\,k\in\mathcal{S}} \tilde{W}_{l,\mathcal{S}}. \quad (15)$$

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 users

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, we have

$$X_{1\to\{2\}} = \tilde{W}_{d_2,\{1\}}\bigcup W^{1\to\{2\}}_{d_2,\{1,3\}}, \quad (16)$$

[1]For convenience, we dropped the subscript $\boldsymbol{d}$ from $X_{j\to\mathcal{T},\boldsymbol{d}}$.

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, i.e., 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

$$X_{1\to\{3\}} = \tilde{W}_{d_3,\{1\}}\bigcup W^{1\to\{3\}}_{d_3,\{1,2\}}, \quad (17)$$

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. On the other hand, 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

$$X_{1\to\{2,3\}} = W^{1\to\{2,3\}}_{d_2,\{1,3\}}\oplus W^{1\to\{2,3\}}_{d_3,\{1,2\}}, \quad (18)$$

where $W^{1\to\{2,3\}}_{d_2,\{1,3\}}\subset\tilde{W}_{d_2,\{1,3\}}$, $W^{1\to\{2,3\}}_{d_3,\{1,2\}}\subset\tilde{W}_{d_3,\{1,2\}}$, and $|X_{1\to\{2,3\}}| = |W^{1\to\{2,3\}}_{d_2,\{1,3\}}| = |W^{1\to\{2,3\}}_{d_3,\{1,2\}}| = v_{1\to\{2,3\}}F$ bits.

From (16)-(18), 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 prevent users $\{1,3\}$ from transmitting redundant bits to user 2 from $\tilde{W}_{d_2,\{1,3\}}$, we need to ensure

$$W^{1\to\{2\}}_{d_2,\{1,3\}}\bigcap W^{1\to\{2,3\}}_{d_2,\{1,3\}}\bigcap W^{3\to\{2\}}_{d_2,\{1,3\}}\bigcap W^{3\to\{1,2\}}_{d_2,\{1,3\}} = \phi. \quad (19)$$

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. That is the transmission and assignment variables represent the size and structure of $X_{j\to\mathcal{T}}$.

First, the structure of the unicast signals in (16) and (17) is represented by the following equality constraints

$$v_{1\to\{2\}} = a_{\{1\}} + u^{1\to\{2\}}_{\{1,3\}},\ v_{1\to\{3\}} = a_{\{1\}} + u^{1\to\{3\}}_{\{1,2\}}. \quad (20)$$

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

$$v_{2\to\{1\}} = a_{\{2\}} + u^{2\to\{1\}}_{\{2,3\}},\ v_{2\to\{3\}} = a_{\{2\}} + u^{2\to\{3\}}_{\{1,2\}}, \quad (21)$$

$$v_{3\to\{1\}} = a_{\{3\}} + u^{3\to\{1\}}_{\{2,3\}},\ v_{3\to\{2\}} = a_{\{3\}} + u^{3\to\{2\}}_{\{1,3\}}. \quad (22)$$

On the other hand, the structure of the multicast signal in (18) is represented by

$$v_{1\to\{2,3\}} = u^{1\to\{2,3\}}_{\{1,3\}} = u^{1\to\{2,3\}}_{\{1,2\}}. \quad (23)$$

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

$$v_{2\to\{1,3\}} = u^{1\to\{2,3\}}_{\{2,3\}} = u^{1\to\{2,3\}}_{\{1,2\}}, \quad (24)$$

$$v_{3\to\{1,2\}} = u^{3\to\{1,2\}}_{\{2,3\}} = u^{3\to\{1,2\}}_{\{1,3\}}. \quad (25)$$

Furthermore, (19) ensures that $\tilde{W}_{d_2,\{1,3\}}$ is divided into disjoint pieces which prevents the transmission of redundant bits. Hence, we have

$$u^{1\to\{2\}}_{\{1,3\}} + u^{1\to\{2,3\}}_{\{1,3\}} + u^{3\to\{2\}}_{\{1,3\}} + u^{3\to\{1,2\}}_{\{1,3\}} \le a_{\{1,3\}}. \quad (26)$$

Similarly, the redundancy constraints for the subfiles $\tilde{W}_{d_3,\{1,2\}}$ and $\tilde{W}_{d_1,\{2,3\}}$, are given by

$$u_{\{1,2\}}^{1\to\{3\}}+u_{\{1,2\}}^{1\to\{2,3\}}+u_{\{1,2\}}^{2\to\{3\}}+u_{\{1,2\}}^{2\to\{1,3\}}\leq a_{\{1,2\}}, \quad (27)$$

$$u_{\{2,3\}}^{2\to\{1\}}+u_{\{2,3\}}^{2\to\{1,3\}}+u_{\{2,3\}}^{3\to\{1\}}+u_{\{2,3\}}^{3\to\{1,2\}}\leq a_{\{2,3\}}. \quad (28)$$

Furthermore, we need to ensure that the transmitted signals complete the requested files, i.e., we have the following delivery completion constraints

$$v_{2\to\{1\}}+v_{2\to\{1,3\}}+v_{3\to\{1\}}+v_{3\to\{1,2\}}\geq 1-\sum_{\mathcal{S}\subset[K]:1\in\mathcal{S}} a_{\mathcal{S}}, \quad (29)$$

$$v_{1\to\{2\}}+v_{1\to\{2,3\}}+v_{3\to\{2\}}+v_{3\to\{1,2\}}\geq 1-\sum_{\mathcal{S}\subset[K]:2\in\mathcal{S}} a_{\mathcal{S}}, \quad (30)$$

$$v_{1\to\{3\}}+v_{1\to\{2,3\}}+v_{2\to\{3\}}+v_{2\to\{1,3\}}\geq 1-\sum_{\mathcal{S}\subset[K]:3\in\mathcal{S}} a_{\mathcal{S}}, \quad (31)$$

since $\sum_{\mathcal{S}\subset[K]:k\in\mathcal{S}} a_{\mathcal{S}}$ is the local caching gain at user k.

Therefore, the set of feasible linear delivery schemes for a three-user system is defined by (20)-(31), and $u_{\mathcal{S}}^{j\to\mathcal{T}}\in[0,a_{\mathcal{S}}]$.

C. Delivery phase: K users

The unicast signal transmitted by user j to user i is

$$X_{j\to\{i\}}=\tilde{W}_{d_i,\{j\}}\bigcup\Big(\bigcup_{\mathcal{S}\subset[K]\setminus\{i\}\,:\,j\in\mathcal{S},|\mathcal{S}|\geq 2} W_{d_i,\mathcal{S}}^{j\to\{i\}}\Big), \quad (32)$$

where $W_{d_i,\mathcal{S}}^{j\to\{i\}}\subset\tilde{W}_{d_i,\mathcal{S}}$ such that $|W_{d_i,\mathcal{S}}^{j\to\{i\}}|=u_{\mathcal{S}}^{j\to\{i\}}F$ 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 $W_{d_i}^{j\to\mathcal{T}}$, 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

$$\mathcal{B}_i^{j\to\mathcal{T}}\triangleq\Big\{\mathcal{S}\subset[K]\setminus\{i\}:\{j\}\cup(\mathcal{T}\setminus\{i\})\subset\mathcal{S}\Big\}, \quad (33)$$

where $\mathcal{B}_i^{j\to\mathcal{T}}$ 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

$$X_{j\to\mathcal{T}}=\oplus_{i\in\mathcal{T}}W_{d_i}^{j\to\mathcal{T}}=\oplus_{i\in\mathcal{T}}\Big(\bigcup_{\mathcal{S}\in\mathcal{B}_i^{j\to\mathcal{T}}} W_{d_i,\mathcal{S}}^{j\to\mathcal{T}}\Big). \quad (34)$$

Furthermore, the set of feasible D2D linear delivery schemes, $\mathfrak{D}(\boldsymbol{a})$, is defined by the following constraints

$$v_{j\to\{i\}}=a_{\{j\}}+\sum_{\mathcal{S}\subset[K]\setminus\{i\}\,:\,j\in\mathcal{S},|\mathcal{S}|\geq 2} u_{\mathcal{S}}^{j\to\{i\}},\forall j\in[K],\,\forall\, i\in\mathcal{T}, \quad (35)$$

$$v_{j\to\mathcal{T}}=\sum_{\mathcal{S}\in\mathcal{B}_i^{j\to\mathcal{T}}} u_{\mathcal{S}}^{j\to\mathcal{T}},\forall j\in[K],\,\forall\,\mathcal{T}\subsetneq_\phi[K]\setminus\{j\},\,\forall i\in\mathcal{T}, \quad (36)$$

$$\sum_{j\in\mathcal{S}}\sum_{\mathcal{T}\subset\{i\}\cup(\mathcal{S}\setminus\{j\})\,:\,i\in\mathcal{T}} u_{\mathcal{S}}^{j\to\mathcal{T}}\leq a_{\mathcal{S}},\,\forall\, i\notin\mathcal{S},\ \forall\,\mathcal{S}\subset[K]\text{ s.t. }2\leq|\mathcal{S}|\leq K-1, \quad (37)$$

$$\sum_{j\in[K]\setminus\{k\}}\sum_{\mathcal{T}\subset[K]\setminus\{j\}\,:\,k\in\mathcal{T}} v_{j\to\mathcal{T}}\geq 1-\sum_{\mathcal{S}\subset[K]:k\in\mathcal{S}} a_{\mathcal{S}},\,\forall\, k\in[K], \quad (38)$$

$$0\leq u_{\mathcal{S}}^{j\to\mathcal{T}}\leq a_{\mathcal{S}},\forall j\in[K],\forall\,\mathcal{T}\subsetneq_\phi[K]\setminus\{j\},\forall\,\mathcal{S}\in\mathcal{B}^{j\to\mathcal{T}}, \quad (39)$$

Algorithm 1 D2D delivery

Input: $\boldsymbol{d},\boldsymbol{a},\boldsymbol{u},\boldsymbol{v}$, and $\tilde{W}_{l,\mathcal{S}}$
Output: $X_{j\to\mathcal{T}}$, $\forall j\in[K]$, $\forall\,\mathcal{T}\subsetneq_\phi[K]\setminus\{j\}$

{Partitioning}
for $\{\mathcal{S}\subset[K]:2\leq|\mathcal{S}|\leq K-1\}$ **do**
  for $\{i\in[K]:i\notin\mathcal{S}\}$ **do**
    Divide $\tilde{W}_{d_i,\mathcal{S}}$ into $W_{d_i,\mathcal{S}}^{j\to\mathcal{T}}$, $\forall j\in\mathcal{S},\forall\,\mathcal{T}\subset\{i\}\cup(\mathcal{S}\setminus\{j\})$ s.t. $i\in\mathcal{T}$, where $|W_{d_i,\mathcal{S}}^{j\to\mathcal{T}}|=u_{\mathcal{S}}^{j\to\mathcal{T}}F$ 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**
      $X_{j\to\{i\}}\leftarrow\tilde{W}_{d_i,\{j\}}\bigcup\Big(\bigcup_{\mathcal{S}\subset[K]\setminus\{i\}\,:\,j\in\mathcal{S},|\mathcal{S}|\geq 2} W_{d_i,\mathcal{S}}^{j\to\{i\}}\Big)$
    else
      $X_{j\to\mathcal{T}}\leftarrow\oplus_{i\in\mathcal{T}}\Big(\bigcup_{\mathcal{S}\in\mathcal{B}_i^{j\to\mathcal{T}}} W_{d_i,\mathcal{S}}^{j\to\mathcal{T}}\Big)$
    end if
  end for
end for

where $\mathcal{B}^{j\to\mathcal{T}}\triangleq\bigcup_{i\in\mathcal{T}}\mathcal{B}_i^{j\to\mathcal{T}}$. Note that (35) follows from the structure of the unicast signals in (32), (36) follows from the structure of the multicast signals in (34), (37) generalizes the redundancy constraints illustrated by (26)-(28) for $K=3$, and (38) generalizes the delivery completion constraints illustrated by (29)-(31) for $K=3$. The delivery procedure is summarized in Algorithm 1.

Remark 2. *The side information constraints were introduced in [9], to ensure decodability at the users. However, decodability can also be guaranteed by combining (35)-(37).* ■

D. Numerical results

Next, we illustrate the solution of (3) by an example, which represents the case in Theorem 4.

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

Placement phase: *Each file $W^{(l)}$ is divided into four subfiles, such that $a_{\{1,2\}}=0.2$, $a_{\{1,3\}}=0.3$, $a_{\{2,3\}}=0.4$, and $a_{\{1,2,3\}}=0.1$.*

Delivery phase: *The D2D transmissions*

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

The caching scheme is illustrated in Fig. 2. Furthermore, $R_{\mathfrak{A}}^*(\boldsymbol{m})=3/2-\big(m_1+m_2+m_3\big)/2=0.45$. Note that the same delivery load is achieved by the caching scheme in [3] for $\boldsymbol{m}=[0.7,\,0.7,\,0.7]$. ■

In Fig. 3, we compare the achievable D2D delivery load $R_{\mathfrak{A},\mathfrak{D}}^*(\boldsymbol{m})$ with the lower bound on $R_{\mathfrak{A}}^*(\boldsymbol{m})$ in Theorem 2,

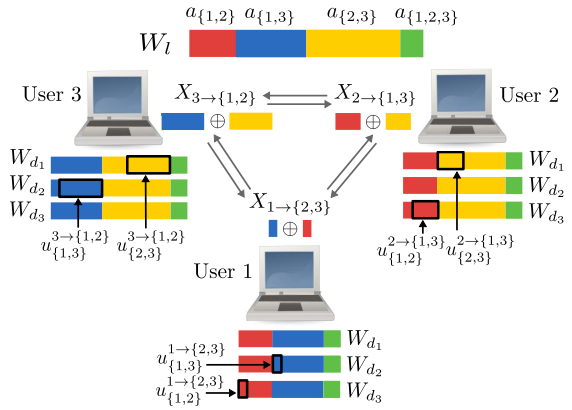

Fig. 2: Example $K = N = 3$, and $\boldsymbol{m} = [0.6,\ 0.7,\ 0.8]$.

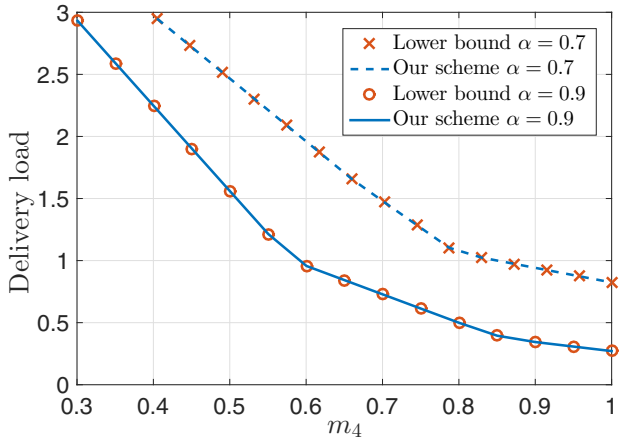

Fig. 3: Comparing $R^*_{\mathfrak{A},\mathfrak{D}}(\boldsymbol{m})$, and lower bound on $R^*_{\mathfrak{A}}(\boldsymbol{m})$, for $K = N = 4$, and $m_k = \alpha\, m_{k+1}$.

for $K = N = 4$ and $m_k = \alpha\, m_{k+1}$. Fig. 3 shows that the lower bound is tight in the general case, i.e., $K > 3$.

VI. Conclusions

In this paper, we have proposed a coded caching scheme for D2D systems when the users are equipped with cache memories of different sizes. We have formulated an optimization problem that characterizes the minimum worst-case D2D delivery load under uncoded placement and linear delivery. Additionally, we have derived a lower bound on the delivery-load memory trade-off under uncoded placement, which enabled us to prove the optimality of our delivery scheme for several cases. In particular, we characterize the trade-off 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) $K = 3$.

Future directions include hierarchical cache-enabled networks and general networks with heterogeneous cache sizes.

Appendix A
Proof of Theorem 2

Under uncoded placement, the D2D-based 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 $|S|$ 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.

Furthermore, by applying the acyclic index-coding bound [12, Corollary 1] on the jth index-coding problem, we get

$$R^{(j)} F \geq \sum_{i=1}^{K-1} \sum_{\mathcal{S}\subset[K]\,:\, j\in\mathcal{S}, \{q_1,\ldots,q_i\}\cap\mathcal{S}=\phi} |\tilde{W}^{(j)}_{d_{q_i},\mathcal{S}}|, \tag{40}$$

where $\boldsymbol{q} \in \mathcal{P}_{[K]\backslash\{j\}}$. Therefore, by considering the convex combinations $\alpha_{\boldsymbol{q}}$ of (40) $\forall \boldsymbol{q} \in \mathcal{P}_{[K]\backslash\{j\}}, \forall j \in [K]$, the minimum over all feasible partitions is given by

$$R^*_{\mathfrak{A}}(\alpha_{\boldsymbol{q}}) \geq \min_{a^{(j)}_{\mathcal{S}}\geq 0} \quad \sum_{j=1}^{K} \tilde{R}^{(j)}(a^{(j)}_{\mathcal{S}}, \alpha_{\boldsymbol{q}}) \tag{41a}$$

$$\text{subject to} \sum_{\mathcal{S}\subsetneq\phi[K]} \sum_{j\in\mathcal{S}} a^{(j)}_{\mathcal{S}} = 1, \tag{41b}$$

$$\sum_{\mathcal{S}\subset[K]:k\in\mathcal{S}} \sum_{j\in\mathcal{S}} a^{(j)}_{\mathcal{S}} \leq m_k, \forall k \in [K], \tag{41c}$$

where $\tilde{R}^{(j)}(a^{(j)}_{\mathcal{S}}, \alpha_{\boldsymbol{q}})$ is defined as

$$(K-1)a^{(j)}_{\{j\}} + \sum_{\substack{\mathcal{S}\subset[K]\,:\, j\in\mathcal{S},\\ 2\leq|\mathcal{S}|\leq K-1}} \left(\sum_{i=1}^{K-|\mathcal{S}|} \sum_{\substack{\boldsymbol{q}\in\mathcal{P}_{[K]\backslash\{j\}}:\, q_{i+1}\in\mathcal{S},\\ \{q_1,\ldots,q_i\}\cap\mathcal{S}=\phi}} i\,\alpha_{\boldsymbol{q}} \right) a^{(j)}_{\mathcal{S}}. \tag{42}$$

By taking the dual of the linear program in (41), and the maximum over all possible convex combinations $\alpha_{\boldsymbol{q}}$, we get the lower bound in Theorem 2. Note that λ_0, and λ_k are the dual variables associated with (41b), and (41c), respectively.

References

[1] Cisco, *Cisco VNI Forecast and Methodology, 2015-2020*, Jun. 2016.

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

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

[4] Z. H. Awan and A. Sezgin, "Fundamental limits of caching in D2D networks with secure delivery," in *Proc. IEEE ICC workshops*, 2015.

[5] A. A. Zewail and A. Yener, "Fundamental limits of secure device-to-device coded caching," in *Proc. IEEE Asilomar*, 2016.

[6] K. Wan, D. Tuninetti, and P. Piantanida, "On the optimality of uncoded cache placement," in *Proc. IEEE ITW*, 2016.

[7] ——, "A novel index coding scheme and its application to coded caching," *arXiv:1702.07265*, 2017.

[8] Q. Yu, M. A. Maddah-Ali, and A. S. Avestimehr, "The exact rate-memory tradeoff for caching with uncoded prefetching," *arXiv:1609.07817*, 2016.

[9] A. M. Ibrahim, A. A. Zewail, and A. Yener, "Centralized coded caching with heterogeneous cache sizes," in *Proc. IEEE WCNC*, 2017.

[10] ——, "Optimization of heterogeneous caching systems with rate limited links," in *Proc. IEEE ICC*, 2017.

[11] 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.

[12] F. Arbabjolfaei, B. Bandemer, Y.-H. Kim, E. Şaşoğlu, and L. Wang, "On the capacity region for index coding," in *Proc. IEEE ISIT*, 2013.