Towards Jointly Optimal Placement and Delivery: To Code or Not to Code in Wireless Caching Networks

Yousef AlHassoun*, Faisal Alotaibi†, Aly El Gamal† and Hesham El Gamal*

*Department of Electrical and Computer Engineering, The Ohio State University, USA.
Email: {alhassoun.1, alotaibi.12, elgamal.2}@osu.edu.
†Department of Electrical and Computer Engineering, Purdue University, USA.
Email: elgamala@purdue.edu

Abstract—Coded caching techniques have received significant attention lately due to their provable gains in reducing the cost of data delivery in wireless networks. These gains, however, have only been demonstrated under the assumption of a free placement phase. This unrealistic assumption poses a significant limitation, especially in cases where aggressive placement strategies can lead to a significant transmission cost that may even be higher than the corresponding cost of the delivery phase. In this paper, we relax this assumption and propose a general caching framework that captures the transmission cost of the two phases, and hence, results in minimizing the overall rate of the caching network. We model the dynamic nature of the network through a cost structure that allows for varying the network architecture and cost per transmission, across the placement and delivery phases. We start with the scenario where the individual users have no limit on the available caching memory and characterize the jointly optimal solution as a function of the different parameters in our cost structure. Then, we characterize the effect of memory constraints on the optimal solution in certain special cases. Interestingly, our results identify regions where the uncoded caching scheme outperforms its coded counterpart. Further, coded caching is shown to offer performance gains only when the network architecture during the placement phase is different from that during the delivery phase.

1. INTRODUCTION

The exponential growth of demands from throughput hungry mobile applications, e.g., mobile video streaming, is causing significant burden on the available wireless infrastructure. This strain on the network is particularly noticeable during peak congestion times where it results in serious degradation in the quality of service offered to end users. Interestingly, several studies have established that during the off peak times the network infrastructure is largely underutilized, as the network is built to handle the peak capacity bandwidth demand [1], [2]. This observation led to the development of various approaches that aim to balance the traffic across both the peak and off peak times.

For example, the cognitive radio paradigm attempts to address this problem by allowing secondary users to access the wireless spectrum during the times where it is underutilized by the primary users, and hence, improving the overall utilization of the available wireless spectrum [3]–[5]. Another example corresponds to content caching and Wi-Fi offloading where certain portions of the requests are rescheduled to happen during off peak times and/or when the users have access to Wi-Fi networks [6]–[8]. This approach can be further improved through the notion of proactive caching where the predictability of user demands and network congestion patterns are used to facilitate a better match between supply, in terms of the available spectrum, and demand in terms of user requests [9]–[12].

Our work here focuses on the coded caching approach, which aims to exploit the multi-casting gain to reduce the cost of data delivery in wireless networks (see e.g., [13]–[15]). In this paradigm, the communication between the source, i.e., Service Provider (SP), and destinations, i.e., end users, is divided into two distinct phases. In the first, referred to as placement phase, the SP transmits information to be cached in the local storage of each end user. This information can be different across users and, it is important to note that, in the current literature the communication cost in the placement phase is assumed to be zero (i.e., free communication). In the second phase, i.e., delivery, by leveraging the shared nature of the wireless channel, coded multi-casting is judiciously applied to minimize the overall cost of delivery (which is typically modeled through the information theoretic rate needed to ensure reliable communication). Of particular relevance to our work is the optimization theoretic framework proposed in [16], [17]. In the sequel, we will adopt several components of this approach in our attempt to understand the impact of the communication cost in the placement phase on the overall efficiency of wireless caching networks.

As alluded to earlier, existing works on coded caching for wireless networks have adopted the notion of cost free communication in the placement phase. This assumption poses, at least, two serious limitations, namely, 1) It inhibits our ability to understand the effect of the varying network architecture and/or cost of transmission, across the placement and delivery phases, on the design of optimal placement and delivery strategies. For example, consider the scenario where the placement phase happens over a Wi-Fi network whereas the delivery phase happens over a cellular network. It is clear that the transmission cost structure in the two phases is very
different and it is natural to inquire on whether this difference will have an impact on the structure of the optimal placement and delivery strategies, and 2) By focusing on the delivery phase only, one may reach misleading conclusions on the overall efficiency of the coded caching approach. To illustrate this point, consider the extreme scenario where the caching in the placement phase ends up being rather aggressive to the extent that the transmission rate in the so called off peak period is actually higher than the optimized multi-cast rate in the delivery phase. Clearly, any conclusions driven from the multi-cast rate in this case is misleading since the network is now limited by the rate required in the placement phase.

This paper attempts to address this problem by introducing a framework for the joint optimization of placement and delivery strategies under more realistic assumptions on the cost of communication. Our framework allows for varying the relative communication cost, between the placement and delivery phases, in two distinct ways. First, our model includes a different scaling factor for the cost per transmission in each phase. Additionally, while we adopt the shared medium assumption in the delivery phase (corresponding to a single cell network), we allow for a different network architecture in the placement phase. More precisely, a single transmission cell network), we allow for a different network architecture in the placement phase. Clearly, any conclusions driven from the multi-cast rate in this case is misleading since the network is now limited by the rate required in the placement phase.

We adopt the objective function of minimizing the maximum transmission cost across the two phases and use this framework to extract valuable insights on the structure of optimal schemes, and to build a more precise characterization of the relative gains that can be leveraged from coded multi-cast caching in realistic settings.

The rest of this paper is organized as follows. In Section II, we layout the system setup and define the characteristics of its main components. We formulate the caching scheme and pose our optimization problem in Section III. In Section IV-A, we a provide a closed form solution with no memory constraints, while we consider memory constraints and provide solutions for special cases in Section IV-B. Numerical results are provided in Section V, and the paper is concluded in Section VI.

II. System Model

Throughout this paper, calligraphic letters (e.g., $\mathcal{X}$) denote finite alphabet sets, boldface letters (e.g., $\mathbf{X}$) denote vectors and non-boldface letters (e.g., $X$) denote scalars. Also, we denote by $\mathbb{N}$ the set of all natural numbers. Moreover, we denote by $|\mathbf{X}|$ the size of vector $\mathbf{X}$.

We consider a service provider (SP) who supplies $N \in \mathbb{N}$ equal sized files to $K \in \mathbb{N}$ users through a shared error-free link, over two phases namely, the placement phase and the delivery phase, where $K \leq N$. Let $\mathcal{N} \triangleq \{1, 2, ..., N\}$ and $\mathcal{K} \triangleq \{1, 2, ..., K\}$ denote the set of file indices and the set of user indices, respectively. We denote the file with index $n \in \mathcal{N}$ by $W_n$, whereas we assume that every user $k \in \mathcal{K}$ has an isolated cache memory of $M \leq N$ files.

Furthermore, in order to transmit to $r \in \mathcal{K}$ users, during the placement phase, the SP incurs a cost $c_r = \rho r^\alpha$, where $\rho \in [0, 1]$ is a linear cost multiplier. When $\alpha = 0$, the cost $c_r$ is equivalent to the cost of a broadcast channel, whereas when $\alpha = 1$, the cost $c_r$ is equivalent to the cost of a time-division multiple access (TDMA) channel. Let $R_\alpha$ and $R_\alpha$ be the rate of transmission that the SP incurs during the placement phase and delivery phase, respectively.

In order to minimize the peak time rate, the SP facilitates caching during the placement phase. In particular, the SP will transmit $Y_o$ over the shared link to all users, and every user $k$ will store a part $Z_k$ in their memories. During the delivery phase, the SP will receive $K$ requests from the users. We will focus in this paper on the worst case scenario, where the SP will receive $K$ different requests in the delivery phase. Let $D_k$ denote the index of the file requested by user $k$ during the delivery phase, and let $D = (D_1, ..., D_K)$ denote the sequence of all user requests. Consequently, the SP broadcasts a multicast message $Y_p$ over the shared link, which will be used for reconstructing the requested message with the aid of the cached content at each user.

III. Problem Formulation

In our formulation, we adopt the caching formulation proposed by [16], which can be generalized as follows:

A. Placement Phase:

The SP partitions each file $W_n$ into $|\mathcal{P}(\mathcal{K})|$ non-overlapping subfiles, where $\mathcal{P}(\mathcal{K})$ is the powerset of the total number of users; these subfiles are denoted by $W_{n,S}$, where $S \in \mathcal{P}(\mathcal{K})$. Moreover, the subfiles are classified according to their types $t \in \{0, 1, 2, ..., K\}$, where $t = |S|$. An example of subfiles for file $W_n$ if we have $K = 3$ would be $W_{n,\{1,2\}}$ with type $t = 2$, and $W_{n,\{1,2,3\}}$ with type $t = 3$. The type will help us to group and label subfiles according to their role. Fig. 1 demonstrates this classification.

![Fig. 1: Subfiles are classified according to their types.](image)

After this division, the SP will broadcast in the placement phase (off-peak time) a concatenated message $Y_o$ of all subfiles of type $t > 0$ for all files to all users, that is:

$$Y_o = \left(W_{n,S} : n \in \mathcal{N}, S \in \mathcal{P}(\mathcal{K}) \setminus \emptyset\right),$$

(1)
where $\emptyset$ is the empty set. Upon receiving this broadcast message, every user $k$ will store subfiles that have its index, that is:

$$ Z_k = \left( W_{n,S} : n \in \mathcal{N}, S \in \mathcal{P}(K) \backslash \emptyset, k \in S \right). $$ (2)

Let $x_{n,S}$ denote the ratio between the number of bits in $W_{n,S}$ and the number of bits in $W_n$. More precisely, we have:

$$ x_{n,S} = \frac{|W_{n,S}|}{|W_n|} \leq 1. $$ (3)

Also, let the file partitioning vector parameter be:

$$ x = \left( x_{n,S} : n \in \mathcal{N}, S \in \mathcal{P}(K) \right). $$ (4)

Furthermore, following the simplification in [16] under the worst case scenario, we will have that all subfiles with the same $t$ have the same size:

$$ x_{n,S} = x_t, \quad \forall n \in \mathcal{N}, S \in \mathcal{P}(K) : |S| = t. $$ (5)

In other words, we will use $x_t$ to denote the length of any subfile of type $t$ for any file $n$. We note that $x$ should satisfy the following constraints:

$$ 0 \leq x_t \leq 1, \quad \forall t \in \{0, 1, \ldots, K\}, $$ (6)$$

$$ \sum_{t=0}^{K} a_t x_t = 1, $$ (7)

$$ \sum_{t=1}^{K} N_t x_t \leq M, $$ (8)

where $a_t = \binom{K}{t}$. Here, (7) represents the file partition constraint, and (8) ensures that the total size of cached contents at every user does not exceed the available memory. We say that $x$ is feasible if it satisfies (6), (7), and (8). The rate in the placement phase is given by:

$$ R_p(x) = N \sum_{t=1}^{K} c_t a_t x_t. $$ (9)

B. Delivery Phase:

Upon receiving the user requests $D$, the SP distinguishes the needed files and the possible opportunity for simultaneously serving multiple requests for global gain using a coded broadcast message. In particular, for each set $S \in \mathcal{P}(K) \backslash \emptyset$, where $|S| = t$, every $t-1$ users in $S$ share a subfile stored in their memory, and it is needed by the remaining user in $S$. More precisely, for any $k \in S$, the subfile $W_{D_k,S\backslash\{k\}}$ is missing at the memory of user $k$, whereas it is present in the memory of any user in $S\backslash\{k\}$. For example, if we have $K = 3$ and user $k = 1$ requests file $D_1$, the subfile $W_{D_1,S(2,3)}$ is missing at user $k = 1$, while it is available at users $k = 2, 3$. The transmitted coded multicasting message in the delivery phase (peak time) can be described as:

$$ Y_p = \{ \oplus_{k \in S} W_{D_k,S\backslash\{k\}} : S \in \mathcal{P}(K) \backslash \emptyset \}, $$ (10)

where $\oplus$ denotes the bitwise XOR operation.

The rate in the delivery phase is given by:

$$ R_p(x) = \sum_{t=0}^{K-1} a_t b_t x_t, $$ (11)

where $b_t = \frac{K-t}{t+1}$.

C. Optimization Problem:

To avoid minimizing one rate at the cost of the other, we introduce a condition on the placement phase rate that guarantees that another peak is not created at the placement time. That is:

$$ N \sum_{t=1}^{K} c_t a_t x_t \leq \sum_{t=0}^{K-1} a_t b_t x_t. $$ (12)

The SP defines its optimization cost function as minimizing the peak time rate while keeping the off peak time rate below it. Then, the SP optimization problem is given by:

$$ \min_x R_p(x) $$ subject to (6), (7), (8), (12). (13) (14)

IV. RESULTS

We first study the case when the memory constraint in (8) is relaxed, and focus on the impact of the communication cost during the placement phase on the optimal scheme.

A. Relaxed Memory Constraint

We have the following characterization of the optimal scheme with no memory constraints.

Theorem 1. For the worst case scenario with no memory constraints, the optimal placement phase will only have at most two types of subfiles, and the optimal caching decision for type $t$ is given by:

$$ x_t^* = \begin{cases} \frac{1}{a_t} \quad & \text{if } P_t < P_{t-1}, \quad \alpha_t \leq \alpha \leq \hat{\alpha}_{t-1}, \\ \frac{1}{P_t} \quad & \text{if } P_t < P_{t-1}, \quad \hat{\alpha}_{t-1} < \alpha < \hat{\alpha}_{t-2}, \\ 0 \quad & \text{otherwise}, \end{cases} $$

where

$$ P_t = \frac{K-t}{t^{2}(t+1)N}, \quad P_{t-1} = \frac{K-t}{t^{2}(t+1)N}, $$

$$ \hat{\alpha}_t = \log_{t^{1/2}} \left( \frac{t+1}{t+2} \right), $$

$$ x_t^* = \begin{cases} \frac{1}{a_t} \quad & \text{if } P_t < P_{t-1}, \quad \alpha_t \leq \alpha \leq \hat{\alpha}_{t-1}, \\ \frac{1}{P_t} \quad & \text{if } P_t < P_{t-1}, \quad \hat{\alpha}_{t-1} < \alpha < \hat{\alpha}_{t-2}, \\ 0 \quad & \text{otherwise}, \end{cases} $$

where

$$ P_t = \frac{K-t}{t^{2}(t+1)N}, \quad \hat{\alpha}_t = \log_{t^{1/2}} \left( \frac{t+1}{t+2} \right), $$

$$ P_{t-1} = \frac{K-t}{t^{2}(t+1)N}, $$

$$ \hat{\alpha}_t = \log_{t^{1/2}} \left( \frac{t+1}{t+2} \right), $$

$$ P_t = \frac{K-t}{t^{2}(t+1)N}. $$

Proof. (Due to the page limitation, we only provide here a sketch of the proof. The full proof is available in [18]) We first note that the optimization problem described in (13) and (14) is a Linear Programming (LP) problem, then the solution exists at one of the extreme points of the constraints. Given that we have only two constraints (7) and (12), the number of basic
feasible solutions in the theory of linear programming can be defined as the number of linearly independent columns of the constraints coefficient matrix. By having only two solutions for the problem, we apply the graphical method to find the closed form solutions with its conditions depending on $\rho$ and $\alpha$.

Remarks.
1) The optimal caching decision requires at most two types, which depends on the value of $\alpha$ and $\rho$. This implies that the structure of the optimal placement and delivery schemes depends largely on the network architecture, which highlights the simplicity of the optimal schemes.
2) Interestingly, uncoded caching, which corresponds to $t^* = K$, is optimal for any $\rho > 0$ if the cost factor $\alpha \leq \frac{1}{1+K}$. For all other values of $\alpha$, coded caching outperforms uncoded caching.
3) As $\alpha$ increases, the value of the optimal type of subfiles decreases. This is expected as increasing $\alpha$ means that the network architecture during the placement phase makes caching at more destinations more expensive, and hence, the optimal structure ends up favoring caching the same information at a smaller number of users.

B. With Memory Constraints

In this part, the limited individual memory constraint of (8) is imposed. We obtain the following result for the case of a broadcast placement channel ($\alpha = 0$).

**Theorem 2.** For the worst case scenario with memory constraints, the optimal caching decision for type $t$ is given by the following when $\alpha = 0$:

$$x_t^* = \begin{cases} 
\frac{1}{\alpha_t}, & \text{if } \rho \leq P_1 \\
\frac{Nt}{K}, & \text{if } P_1 < M < \frac{N(t+1)}{K}, \\
\frac{N(t-1)}{K}, & \text{if } \rho \leq P_2, \\
\frac{Nt}{K}, & \text{if } \frac{N(t-1)}{K} < M < \frac{Nt}{K}, \\
\frac{Nt}{K}, & \text{if } P_2 < M < M_1, \\
\frac{Nt}{K}, & \text{if } P_3 < M < M_2, \\
\frac{Nt}{K}, & \text{if } P_4 < M < M_3, \\
\frac{Nt}{K}, & \text{if } P_4 \leq M < M_4.
\end{cases}$$

where $P_1 = \frac{Nt + \frac{N}{K} - MK}{t+1}$, $M_t = \frac{N(t+1)}{t+1}$, $P_2 = \frac{Nt}{K-Nt+MK}$, $P_3 = \frac{Nt}{K-Nt+MK}$, $P_4 = \frac{Nt}{K-Nt+MK}$, and $P_4 = \frac{Nt}{K-Nt+MK}$.

Proof. (Sketch) When $\alpha = 0$, the proof follows the same footsteps used in the proof Theorem 1 after using an extra constraint to incorporate the memory limitation, resulting in an LP problem which has a solution at one of the extreme points. By having only three solutions for the problem, we apply the graphical method to find the closed form solutions with its conditions depending on $\rho$ and $\alpha$. It is noteworthy to point out that (7) and (8) are the active constraints used to obtain $P_1$ and $P_2$, while (8) and (12) are the active constraints used to obtain $P_3$ and $P_4$.

**Corollary 1.** The memory size needed to enable the optimal caching decision for type $t$, that is characterized in Theorem 2 is:

$$M = \frac{Nt}{K} \text{ if } \rho = \frac{K-t}{(t+1)N}. \quad (15)$$

Proof. (Sketch) It follows from Theorem 2 that with no memory constraints, the optimal size for subfiles of type $t$ is $x_t^* = \frac{1}{\alpha_t}$ when $\rho = P_1$; this condition implies that the available memory needed from constraint (8) at each user is $\frac{Nt}{K}$, which imposes a restriction towards reaching the optimal type, and hence, the corollary statement follows.

**Corollary 2.** For the worst case scenario with memory constraints, the uncoded caching scheme is optimal if the following two conditions hold:

1) $\alpha \leq \frac{1}{1+K}$,
2) $M \geq \frac{NK}{(\rho NK^\alpha + K)}$.

Proof. (Sketch) It follows from Theorem 1 that with no memory constraints, Condition 1 implies that $t^* = K$. Further, the size of subfiles with type $K$ will be given by,

$$x_K = \frac{K}{(\rho NK^\alpha + K)}. \quad (16)$$

Now, Condition 2 implies that the availability of the memory at each user does not impose a restriction towards reaching the optimal solution, and hence, the theorem statement follows.

Note that $\rho = 0$ corresponds to the scenario with no placement cost. Hence, in this setting, we only have the memory constraint, and the optimal caching type is given by the following corollary of Theorem 2.

**Corollary 3.** When $\rho = 0$, the optimal caching decision obtained in Theorem 2 for type $t$ is given by:

$$x_t^* = \begin{cases} 
\frac{1}{\alpha_t}, & \text{if } t = \frac{MK}{N}, \\
\frac{Nt}{N-MK}, & \text{if } t = \frac{MK}{N}, \\
\frac{Nt}{N-MK}, & \text{if } t = \frac{MK}{N}, \\
0, & \text{otherwise}.
\end{cases}$$

It is worth mentioning that the above corollary is in alignment with the solutions of [13], [16] as they assume no placement cost, i.e., $\rho = 0$, which makes our formulation a general framework for coded caching that takes the network architecture and cost of time resource under consideration.
V. NUMERICAL RESULTS

We provide in this section numerical results for a system consisting of $K = 10$ users interested in a library of $N = 50$ files. From the results in Section IV-A, we plot in Fig. 2 the cost $R_p(x)$ versus the selected type of subfiles for different values of the cost parameter $\alpha$ (i.e., different network architectures during the placement phase). Note that the cost is computed at every value of $t$, not just integer points, as an intermediary step towards reaching the results in Theorem 1 included relaxing the type integer constraint to a continuous variable (see [18] for details).

Interestingly, we established that coded multicast provided a sharp characterization of the optimal placement and delivery: To code or not to code in wireless caching networks,” arXiv preprint arXiv, 2019.

Within this framework, we characterized the optimal scheme along with the possibility of a dynamic network architecture. Cost of transmission, across the placement and delivery phases, networks. Our framework allows for capturing the different optimization of placement and delivery in wireless caching networks.

Fig. 2: Effect of the placement phase network architecture on the optimal caching type.

As $\alpha$ increases, the value of the optimal type of subfiles decreases. In particular, when $\alpha = 0$, the cost of a single transmission during the placement phase does not depend on the number of destinations, which makes it optimal to send only subfiles with type $t = K$ (i.e., all users cache every transmitted bit). Hence, there will be no advantage to be reaped from coded multi-cast in the delivery phase. In other words, if the network has the same architecture during both placement and delivery phases, then uncoded caching is optimal for any valid value of $\rho$.

VI. CONCLUSION

In this work, we introduced a framework for the joint optimization of placement and delivery in wireless caching networks. Our framework allows for capturing the different cost of transmission, across the placement and delivery phases, along with the possibility of a dynamic network architecture. Within this framework, we characterized the optimal scheme that minimizes the transmission cost over the two phases under a worst case assumption for the demands. Furthermore, our analysis yielded a sharp characterization of the optimal type of caching in the placement phase and the scenarios where coded multicast outperforms the uncoded counterpart. Interestingly, we established that coded multicast provided gains only when the network architecture during the placement phase is different from that during the delivery phase (i.e., $\alpha \neq 0$ in our model). Finally, our numerical results further elucidated the dependencies of the achieved cost on the different parameters of the system. Our framework opens the door for further studies on the multi-faceted problem of wireless caching networks’ design. Our current work aims to incorporate statistical objective functions (instead of the worst case cost function), investigate rigorously the impact of other system constraints (e.g., energy) on the design and analysis of optimal caching schemes, and relax some of our modeling assumptions (e.g., the equal file/subfile sizes).

REFERENCES

[1] P. Kolodzy and I. Avoidance, “Spectrum policy task force,” Federal Commun. Comm., vol. 40, no. 4, pp. 147–158, 2002.
[2] J. S. Adelstein et al., “Facilitating opportunities for flexible, efficient, and reliable spectrum use employing spectrum agile radio technologies (et docket no. 03 108),” Washington, DC. USA: FCC.
[3] J. Mitola, “Cognitive radio—an integrated agent architecture for software defined radio,” Royal Institute of Technology (KTH), 2000.
[4] S. Haykin et al., “Cognitive radio: brain-empowered wireless communications,” IEEE journal on selected areas in communications, vol. 23, no. 2, pp. 201–220, 2005.
[5] D. Niyato and E. Hossain, “Competitive pricing for spectrum sharing in cognitive radio networks: Dynamic game, inefficiency of Nash equilibrium, and collusion,” IEEE journal on selected areas in communications, vol. 26, no. 1, 2008.
[6] L. Gao, G. Isostidis, J. Huang, and L. Tassiulas, “Economics of mobile data offloading,” in INFOCOM, 2013 Proceedings IEEE, pp. 3303–3308, IEEE, 2013.
[7] L. Xiaofeng, H. Pan, and P. Lio, “Offloading mobile data from cellular networks through peer-to-peer wifi communication: A subscribe-and-send architecture,” China Communications, vol. 10, no. 6, pp. 35–46, 2013.
[8] J. Lee, Y. Yi, S. Chong, and Y. Jin, “Economics of wifi offloading: Trading delay for cellular capacity,” IEEE Transactions on Wireless Communications, vol. 13, no. 3, pp. 1540–1554, 2014.
[9] H. El Gamal, J. Tadrous, and A. Eryilmaz, “Proactive resource allocation: Tuning predictable behavior into spectral gain,” arXiv preprint arXiv:1010.1800, 2010.
[10] J. Tadrous, A. Eryilmaz, and H. El Gamal, “Proactive resource allocation: Harnessing the diversity and multicast gains,” IEEE Transactions on Information Theory, vol. 59, no. 8, pp. 4833–4854, 2013.
[11] J. Tadrous, A. Eryilmaz, and H. El Gamal, “Pricing for demand shaping and proactive download in smart data networks,” in INFOCOM, 2013 Proceedings IEEE, pp. 3189–3194, IEEE, 2013.
[12] J. Tadrous, A. Eryilmaz, and H. El Gamal, “Proactive content download and user demand shaping for data networks,” IEEE/ACM Transactions on Networking (TON), vol. 23, no. 6, pp. 1917–1930, 2015.
[13] M. A. Maddah-Ali and U. Niesen, “Fundamental limits of caching,” IEEE Transactions on Information Theory, vol. 60, no. 5, pp. 2856–2867, 2014.
[14] M. A. Maddah-Ali and U. Niesen, “Decentralized Coded Caching Attains Order-Optimal Memory-Rate Tradeoff,” IEEE/ACM Transactions on Networking, 2015.
[15] Q. Yu, M. A. Maddah-Ali, and A. S. Avestimehr, “The exact rate-memory tradeoff for caching with uncoded prefetching,” IEEE Transactions on Information Theory, vol. 64, no. 2, pp. 1261–1296, 2018.
[16] S. Jin, Y. Cui, H. Liu, and G. Caire, “Structural properties of uncoded placement optimization for coded delivery,” arXiv preprint arXiv:1707.07146, 2017.
[17] A. M. Daniel and W. Yu, “Optimization of heterogeneous coded caching,” arXiv preprint arXiv:1708.04322, 2017.
[18] Y. AlHassoun, F. Alotaibi, A. El Gamal, and H. El Gamal, “Towards jointly optimal placement and delivery: To code or not to code in wireless caching networks,” arXiv preprint arXiv, 2019.