Coded Caching in Fog-RAN: $b$-Matching Approach

Bo Bai, Senior Member, IEEE, Wanyi Li, Student Member, IEEE, Li Wang, Senior Member, IEEE, Gong Zhang

Abstract

Fog radio access network (Fog-RAN), which pushes the caching and computing capabilities to the network edge, is capable of efficiently delivering contents to users by using carefully designed caching placement and content replacement algorithms. In this paper, the transmission scheme design and coding parameter optimization will be considered for coded caching in Fog-RAN, where the reliability of content delivery, i.e., content outage probability, is used as the performance metric. The problem will be formulated as a complicated multi-objective probabilistic combinatorial optimization. A novel maximum $b$-matching approach will then be proposed to obtain the Pareto optimal solution with fairness constraint. Based on the fast message passing approach, a distributed algorithm with a low memory usage of $O(M + N)$ is also proposed, where $M$ is the number of users and $N$ is the number of Fog-APs. Although it is usually very difficult to derive the closed-form formulas for the optimal solution, the approximation formulas of the content outage probability will also be obtained as a function of coding parameters. The asymptotic optimal coding parameters can then be obtained by defining and deriving the outage exponent region (OER) and diversity-multiplexing region (DMR). Simulation results will illustrate the accuracy of the theoretical derivations, and verify the outage performance of the proposed approach. Therefore, this paper not only proposes a practical distributed Fog-AP selection algorithm for coded caching, but also provides a systematic way to evaluate and optimize the performance of Fog-RANs.

Index Terms

Coded caching, Fog-RAN, $b$-matching, fast message passing, saddle-point method, outage exponent region.

The corresponding authors of this paper are B. Bai (baibo8@huawei.com) and L. Wang (liwang@bupt.edu.cn).

B. Bai and G. Zhang are with Future Network Theory Lab, 2012 Labs, Huawei Technologies Co., Ltd., Shatin, N. T., Hong Kong (e-mail: baibo8@huawei.com, nicholas.zhang@huawei.com).

W. Li and L. Wang are with School of Electronic Engineering, Beijing University of Posts and Telecommunications, Beijing 100876, China (e-mail: lwy@bupt.edu.cn, liwang@bupt.edu.cn).
I. INTRODUCTION

Fog-computing, also known as mobile edge computing (MEC), is a novel and promising technology for future networks [1]. In contrast to conventional base station (BS) and access point (AP), Fog-AP is capable working as a wireless AP, and provides caching and computing capabilities [2]. The network composed by Fog-APs is generally referred to as the fog radio access network (Fog-RAN), which can improve the content delivery efficiency and support computation offloading [3]. Therefore, the problem of caching scheme design in Fog-RAN attracted much attention from both industry and academia.

By pushing contents to the edge, the users can access the interested information within one hop, which significantly reduces the latency. In [4], an information-theoretic formulation of the caching problem is introduced. The basic structure and global caching gain is analyzed for a novel coded caching approach. In [5], the authors focused on content sharing among smart devices in the social IoT with D2D-based cooperative coded caching. In [6], the un-coded and coded caching schemes are studied for femtocells. The fountain code is considered, where the objective is to solve the cache assignment problem with a given maximum number of helpers a user can be connected to. In [7], the MDS code with random caching strategy is considered in D2D networks, where the fundamental limits of caching is obtained. In [8], the authors investigated the content delivery network and information centric network solution of 5G mobile communication network. In [9], both centralized algorithm and distributed algorithm are proposed for optimal caching placement problem in Fog-RAN. In [10], a caching based socially aware D2D communication framework is considered, where a hypergraph framework is summarized.

Previous works illustrate great potential of edge caching for reducing the delay by pushing the user required content at network edge nodes. In this paper, we focus on the transmission scheme design and coding parameter optimization for coded caching in Fog-RAN [11]. As the Fog-AP may encounter device failure, it is required to recover the stored content from other Fog-APs to a new Fog-AP. By using the coded caching scheme, the data can be recovered by transmitting data from other Fog-APs through backhaul link. Thus, the user can still access the interested content in one hop. On the other hand, the wireless channel may be in deep fading, some of the Fog-APs will be in outage in this case. By using coded caching, the user can recover the interested content by accessing any given number of Fog-APs. In this context, the reliability can be greatly improved if the coded caching scheme is applied. In this work, the objective is to choose the optimal subset of Fog-APs for each user, so that the required
content can be delivered to each user with the highest successful probability, i.e., the content outage probability is minimized. Therefore, we need to choose a coding scheme to minimize the size of data needed to be transmitted from Fog-APs to each user, i.e., the coding scheme with minimum storage size. On the other hand, the Fog-APs are assumed to be connected by reliable (wired or wireless) back-haul links with high bandwidth. The bandwidth requirement for regenerating the storing data in a new Fog-AP can be easily guaranteed. In this context, we choose the minimum storage regenerating (MSR) code with the optimal coding parameters. The transmission scheme design and coding parameter optimization problem is a complicated multi-objective probabilistic combinatorial optimization [12]. Moreover, the requirement on solving this problem in a distributed manner further increases the difficulties.

It is well known there is (generally) no global optimum to multi-objective optimization problem [13]. Therefore, we focus on the Pareto optimal solution (i.e., none of the objectives can be improved without degrading other objectives) with fairness constraint for coded caching problem in Fog-RAN. The key is the novel $b$-matching approach, which can solve this problem with fixed coding parameters. Here $b$ is a function defined on vertices, which determines the maximum number of edges associated with a vertex in the matching [14]. Specifically, the Fog-AP selection problem is first formulated as a random bipartite graph (RBG) based fairness maximum $b$-matching problem. To explore not only the caching capability but also computing capability of Fog-APs, a fast message passing algorithm is proposed to select Fog-APs for each user in a distributed way. The memory usage of the
The proposed algorithm is only $O(M + N)$, where $M$ is the number of users and $N$ is the number of Fog-APs. Although, it is very difficult to derive the closed-form formulas for the optimal solution, a tight upper bound for the solution is still obtained by digging the properties of fairness maximum $b$-matching on RBG and conditional content outage probability. To obtain the optimal coding parameters, we define and derive the outage exponent region (OER), the region that contains all feasible outage exponent vectors of all users, and diversity-multiplexing region (DMR), the OER as the signal to noise ratio (SNR) tends to infinity, from content outage probability\footnote{The diversity-multiplexing tradeoff (DMT), which is first proposed in \cite{15}, is used as a key performance metric in high SNR regime. Specifically, the DMT is the slope of the outage probability curve in the high SNR regime. According to the definition in \cite{15}, the DMT is the tradeoff between reliability (diversity gain) and efficiency (multiplexing gain) when SNR tends to infinity. The DMR is the DMT in multiple user scenario. The OER is a generalization of DMR in finite SNRs with multiple users.}. The obtained theoretical results illustrate that $b$-matching approach is optimal in high SNR regime. Specifically, each user not only fairly shares the multiplexing gain but also achieves the full diversity gain, i.e., the largest area of DMR. Simulation results illustrate the accuracy of the theoretical derivations, and verify the outage performance of the proposed scheme. In general, the contribution our work can be summarized as follows:

1) The optimal content transmission scheme from Fog-APs and users is proposed, which is formulated as a multi-objective probabilistic combinatorial optimization problem. The novel $b$-matching approach is proposed for solving this problem. The decentralized algorithm is also obtained to select the optimal subset of Fog-APs for each user with low complexity and small storage cost.

2) The closed-form approximation formulas of the content outage probability is derived in this work. The DMR can be obtained to verify that the solution is optimal in high SNRs. These formulas provide a systematic way to design and evaluate the caching scheme for Fog-RANs in an analytic way.

The rest of this paper is organized as follows. Section II presents the system model and problem formulation. The RBG based $b$-matching approach is presented in Section III where the fast message passing based distributed Fog-AP selection algorithm is discussed in detail. Section IV defines and derives the closed-form formulas for content outage probabilities, OER, and DMR to obtain the optimal coding parameters. Section V presents the simulation results. Finally, Section VI concludes this paper.
II. SYSTEM MODEL AND PROBLEM FORMULATION

A. Channel Model

Consider a dense Fog-RAN, as shown in Fig. 1, where $M$ users communicate with $N$ Fog-APs through orthogonal subchannels. The user set is denoted by $\mathcal{U} = \{u_m\}_{m=1}^{M}$, where $u_m$ is the $m$-th user. The Fog-AP set is denoted by $\mathcal{A} = \{a_n\}_{n=1}^{N}$, where $a_n$ is the $n$-th Fog-AP. One Fog-AP has a dedicated subchannel, each of which contains $L$ resource blocks (RBs) in the duration of the coherence time. For convenience, both the bandwidth and the duration of one RB are normalized as one. In this context, one Fog-AP can be accessed by $L$ users at most, each of which occupies one RB of the subchannel.

Suppose that the signal on each subchannel undergoes the independent Rayleigh slow fading. Then, the channel gain between user $u_m$ and Fog-AP $a_n$, denoted by $h_{mn}$, is independent and follows the identical distribution of $CN(0, 1)$. The received signal from Fog-AP $a_n$ to user $u_m$, denoted by $y_m$, is given by

$$y_m = h_{mn}x_n + w_m,$$

where $w_m \sim CN(0, 1)$ is the additive Gaussian white noise (AWGN). As we assume one bit channel state information (CSI) is known at the transmitter, the mutual information between user $u_m$ and Fog-AP $a_n$ during one RB is given by

$$I_{mn} = \max I(y_m; x_n) = \ln \left( 1 + |h_{mn}|^2 \gamma \right) \text{ nats},$$

(1)

where $\gamma$ is the average SNR at the receiver. Throughout the paper, the unit of information is “nat”. As we normalize both the bandwidth and the duration of one RB as one, the transmission rate is equal to $I_{mn}$ nats per RB use.

B. Coded Caching in Fog-RANs

To fully exploit the caching capability in Fog-APs, the coded caching scheme is adopted. Suppose the user $u_m$ requires the $m$-th content $s_m$, which composes the content set $\mathcal{S} = \{s_m\}_{m=1}^{M}$. The size of content $s_m$ is $R_m$ nats. Thus, there are $M$ files cached in Fog-RAN. As proposed in [11], the distributed storage code with parameters $(N, K_m, D_m)$ is used for content $s_m$. Thus, $s_m$ will be cached in $N$ Fog-APs, each of which stores $\alpha_m$ nats. Hence, the storage size of each Fog-AP is at least $\sum_{m=1}^{M} \alpha_m$ nats. If $K_m$ Fog-APs transmit $\alpha_m$ nats each

3The script symbol $\mathcal{X}$ denotes a set, whose cardinality is denoted by $|\mathcal{X}|$. The abbreviation $\{x_m\}_{m=1}^{M}$ is used to denote $\{x_1, x_2, \ldots, x_M\}$.

\[CN(\mu, \sigma^2)\] denotes a circularly symmetric Gaussian distribution with mean $\mu$ and variance $\sigma^2$. 


to user \( u_m \), it is capable of recovering the original content \( s_m \). For a new Fog-AP without caching any part of \( s_m \), it requires \( D_m \) Fog-APs to transmit \( \beta_m \) nats in total to regenerate \( \alpha_m \) nats in this new Fog-AP. It can be seen that the performance of the distributed storage code is characterized by \((\alpha_m, \beta_m)\) [11]. In this context, we have \( N \geq K_m \) for \( m = 1, 2, \ldots, M \) and \( \sum_{m=1}^{M} K_m \leq NL \).

In the considered Fog-RAN, the contents should be first cached in Fog-APs based on some caching strategies [4], [9]. Then, the method in this work will be applied to optimize the transmission performance for the cached contents. Specifically, Fog-APs are connected by reliable (wired or wireless) back-haul links with high bandwidth [3]. Thus, compared to the wireless channel between users and Fog-APs, the total regenerating bandwidth \( \beta_m \) can be naturally assumed to be fulfilled.\(^4\) In contrast, due to channel fading, the mutual information between user \( u_m \) and Fog-AP \( a_n \) during one RB may be smaller than \( \alpha_m \). In other words, the Fog-AP \( a_n \) is in outage for user \( u_m \), referred to as Fog-AP outage. Thus, we need to minimize \( \alpha_m \), i.e., apply MSR codes [11]. Accordingly, we have the following optimal \((\alpha_m^*, \beta_m^*)\) for MSR codes:

\[
(\alpha_m^*, \beta_m^*) = \left( R_m \frac{D_m}{K_m}, \frac{D_m}{D_m - K_m + 1} \frac{R_m}{K_m} \right)
\]  \( (2) \)

C. Fog-AP Selection Problem

According to the property of MSR codes, the user must successfully access \( K_m \) Fog-APs in order to recover the required content \( s_m \). Due to channel fading, however, the Fog-AP \( a_n \) will be in outage for user \( u_m \), if \( I_{mn} < \alpha_m^* \). In this context, the user may still fail to recover \( s_m \), even the MSR code with \((\alpha_m^*, \beta_m^*)\) is applied, i.e., the user may encounter content outage. Intuitively, with the fixed transmission rate \( \alpha_m^* \) from each Fog-AP to user \( u_m \) in one RB, the content outage probability decreases if we decrease \( K_m \). However, as \( \alpha_m^* \) is equal to \( \frac{R_m}{K_m} \), the Fog-AP outage probability increases if we decrease \( K_m \). Therefore, an optimal \( K_m \) exists for achieving the minimum content outage probability. On the other hand, because of the flexible architecture of the Fog-RAN, several Fog-APs are able to form an edge cloud to support mobile edge computing. Thus, the joint channel coding scheme, such as rotated \( Z^{K_m} \)-lattice code [16] or permutation code [17], can be applied in Fog-APs so as to further

\(^{4}\)If we take back-haul links into consideration, there will be a tradeoff between \( \alpha_m \) and \( \beta_m \). The proposed framework in this paper can be easily extended to this scenario.
improve the outage performance. According to [16]–[18], the outage probability of user $u_m$ for content $s_m$ is given by

$$p_{out}^m(R_m) = \Pr \left\{ \sum_{a_n \in A_m} I_{mn} < R_m \right\}, \quad (3)$$

where $R_m = K_m \alpha_m^*$, and $A_m$ is the set of Fog-APs accessible by user $u_m$, which satisfies

$$|A_m| = K_m; \quad \bigcup_{m=1}^{M} A_m \subseteq A. \quad (4)$$

To reduce the complexity in both signaling and computation, each user is only allowed to feedback 1 bit CSI for each subchannel to Fog-APs at the beginning of each coherence time, i.e., we only knows whether the Fog-AP $a_n$ is in outage for user $u_m$ or not. Let $Q$ denote the 1 bit quantized CSI matrix, the entry at the $m$-th row and $n$-th column is given by

$$[Q]_{mn} = \begin{cases} 0, & I_{mn} < \alpha_m^*; \\ 1, & I_{mn} \geq \alpha_m^*. \end{cases} \quad (5)$$

Therefore, the Fog-AP outage probability is given by

$$p_m = \Pr \{[Q]_{mn} = 0\} = 1 - \exp \left( -\frac{\alpha_m^* - 1}{\gamma} \right). \quad (6)$$

For convenience, we define

$$q_m = \Pr \{[Q]_{mn} = 1\} = 1 - p_m = \exp \left( -\frac{\alpha_m^* - 1}{\gamma} \right). \quad (7)$$

In this context, a Fog-AP selection scheme can be seen as a mapping from $Q$ to $\{A_m\}_{m=1}^{M}$, that is

$$\mathcal{S} : Q \rightarrow \{A_m\}_{m=1}^{M}. \quad (8)$$

According to the notation in [13] and footnote 1 of this paper, the optimal Fog-AP selection problem can be formulated as follows:

$$(P1) \quad \min_{\mathcal{S}} \quad \{p_{out}^m(R_m)\}_{m=1}^{M}$$

s.t. $$\sum_{m=1}^{M} K_m \leq NL$$

$$K_m < N. \quad (9)$$

The problem P1 in Eq. (9) summarizes the optimal Fog-AP selection problem in an abstraction form. It will determine the optimal mapping $\mathcal{S}$ and $\{K_m\}_{m=1}^{M}$ so as to minimize the content outage probability in Eq. (3). As a multi-objective probabilistic combinatorial optimization
problem, P1 has no global optimum in general according to [12], [13]. Therefore, the Pareto optimal is proposed as a solution for multi-objective optimization problem. Specifically, the strong Pareto boundary of a multi-objective optimization problem consists of the attainable operating points that none of the objectives can be improved without degrading other objectives. Every point on the strong Pareto boundary is a Pareto optimal solution. In this work, we focus on the Pareto optimal solution with fairness consideration, denoted by \( S^* \) and \( \{ K_m \}^M_{m=1} \) \(^5\). In the following, with fixed \( K_m \) P1 is first reformulated and solved as P2 by the proposed \( b \)-matching approach in Section [III]. The \( b \)-matching approach guarantees that all of the users fairly share the total channel resource. Specifically, each user will be allocated the channel resource which is proportional to its rate requirement. In Section [IV] the content outage probability, as a function of \( K_m \), is obtained in closed-form formulas, so that \( K_m \) can be optimized. The theoretical results show that the \( b \)-matching approach also achieves the optimal diversity order, i.e., the total diversity gain provided by the channel. Therefore, the Pareto optimal solution with fairness constraint is obtained by the proposed \( b \)-matching approach.

III. RBG BASED \( b \)-MATCHING APPROACH

In this section, P1 with fixed \( K_m \) is first reformulated as a combinatorial optimization problem P2 based on a given sample of \( Q \). Then, the RBG based \( b \)-matching method is proposed for solving P2, which is exactly the Pareto optimal mapping \( S^* \) with fairness consideration for P1. To fully exploit the computation capability of Fog-APs, a fast message passing based distributed Fog-AP selection algorithm is proposed.

A. Problem Re-formulation

At the beginning epoch of each coherence time, \( Q \) can be obtained by channel estimation. The optimal Fog-AP selection problem will then be solved based on the known \( Q \). Define the decision variable \( x_{mn} \) as

\[
x_{mn} = \begin{cases} 
1, & \text{if } a_n \in \mathcal{A}_m; \\
0, & \text{otherwise.} 
\end{cases}
\]

(10)
The decision matrix with \( x_{mn} \) being the entry at the \( m \)-th row and \( n \)-th column is denoted by \( X \). With the fixed \( K_m \) and known \( Q \), the mapping \( S \) can be represented by the decision

\(^5\)According to [13], the scalarization of this problem can be obtained by specifying a certain subjective tradeoff between the objectives. In this work, the Pareto optimal with fairness constraint is used as the scalarization method. As an another scalarization of the problem, the min-max formulation will be investigated in our future work.
matrix $X$. To achieve the Pareto optimal solution, i.e., minimum content outage probability, the objective is to maximize the total number of non-outage Fog-APs which are accessed by all of the users. The fairness constraint means that $u_m$ will access $\eta K_m$ non-outage Fog-APs has the same probability $\epsilon$, where $\eta < 1$ and $\epsilon < 1$. In this context, the fair Pareto optimal mapping $S^*$ is equivalent to solve the following combinatorial optimization problem:

$$\max_X \sum_{m=1}^{M} \sum_{n=1}^{N} [Q]_{mn} x_{mn}$$

s.t. $\sum_{m=1}^{M} x_{mn} \leq L$, $\sum_{n=1}^{N} x_{mn} = K_m$, $\Pr \left\{ \frac{\sum_{n=1}^{N} [Q]_{mn} x_{mn}}{K_m} = \eta \right\} = \epsilon$, $x_{mn} \in \{0, 1\}$, $m = 1, \ldots, M$, $n = 1, \ldots, N$.

It can be seen that the content outage probability in P1 is minimized when combining the optimal solution of P2 and joint coding scheme for every sample of $Q$. The first constraint implies that each Fog-AP can only support $L$ users, each of which occupies one RB. The second constraint indicates that each user will access $K_m$ Fog-APs to fulfill the requirement of MSR codes. Due to channel fading, however, the accessible $K_m$ Fog-APs for user $u_m$ may not always be non-outage for every sample of $Q$. Therefore, the third constraint is introduced to guarantee fairness.

B. $b$-Matching based Fog-AP Selection

P2 is a complicated combinatorial optimization problem. The RBG based $b$-matching method is proposed in the following to solve P2, which is also the Pareto optimal mapping $S^*$ with fairness consideration for P1. The preliminaries on RBG and $b$-matching can be found in Appendix A.

The RBG model for Fog-RAN can be constructed as follows. All of the vertices in one partition class represent all of the users in $\mathcal{U}$, and all of the vertices in the other partition class represent all of the Fog-APs in $\mathcal{A}$. If $[Q]_{mn} = 1$, we join the vertex $u_m \in \mathcal{U}$ and the vertex $a_n \in \mathcal{A}$ with an edge $e = (u_m, a_n)$. Otherwise, there will be no edge between $u_m$ and $a_n$. Therefore, the probability space of the RBG model can be denoted by $\mathcal{G}\{K_{MN}; P\}$ with $P = \{p_m\}_{m=1}^{M}$, where $p_m$ is given by Eq. (6). A sample of $\mathcal{G}\{K_{4,3}; P\}$ is shown in Fig. 2.
Fig. 2. A sample of $G(K_m; P)$ with $K_m = 2$ and $L = 3$. The thick segments are the edges in $M_b^{m}$. The dashed segment denotes the selected outage Fog-AP.

It can be seen that each sample of $G(K_M; P)$, denoted by a bipartite graph $G(U \cup A, \mathcal{E})$, corresponds to a snapshot of Fog-RAN and vice versa. For any given $G(U \cup A, \mathcal{E})$, associate $b(v)$ to $v \in U \cup A$ as

$$b(v) = \begin{cases} K_m, & v = u_m \in U; \\ L, & v = a_n \in A. \end{cases}$$  (12)

As the maximum $b$-matching $M_b^{m}$ may not be unique on some samples, to guarantee the third constraint in problem P2, we need to generate $M_b^{m}$ on $G(U \cup A, \mathcal{E})$ with $\Pr \left\{ \frac{k_m}{K_m} = \eta \right\} = \epsilon$ for every $u_m \in U$, where $k_m = |M_b^{m} \cap \mathcal{E}(u_m)|$ is the number of non-outage Fog-APs allocated to $u_m$. The maximum $b$-matching satisfying the fairness constraint is referred to as the fairness maximum $b$-matching and denoted by $M_b^{m}$. In this context, the Fog-AP set $A_m^*$ selected for user $u_m$ can be constructed as

$$A_m^* = A(M_b^{m} \cap \mathcal{E}(u_m)) \cup \tilde{A}_m,$$  (13)

where $A(M_b^{m} \cap \mathcal{E}(u_m)) \subseteq A$ contains all of the vertices that are incident with an edge in $M_b \cap \mathcal{E}(u_m)$, $\tilde{A}_m \subseteq A$ contains $K_m - k_m$ Fog-APs, each of which satisfies

$$\sum_{m'=1}^M 1_{A_m^{m'}}(a_n) \leq L, \quad \forall a_n \in \tilde{A}_m,$$  (14)

where $1_{\mathcal{X}}(x)$ is the indicator function defined as

$$1_{\mathcal{X}}(x) = \begin{cases} 1, & x \in \mathcal{X}; \\ 0, & x \notin \mathcal{X}. \end{cases}$$  (15)
Clearly, if $\mathcal{M}_{b}^{fm} = \emptyset$, $\bar{A}_m$ is an empty set for any $u_m \in \mathcal{U}$. Otherwise, $\bar{A}_m$ can be generated by randomly selecting $K_m - k_m$ non-saturated outage Fog-APs to user $u_m$. The discussion above implies that the optimal solution of $P_2$ is given by:

$$x_{mn}^* = \begin{cases} 1, & \text{if } a_n \in \mathcal{A}_m^*; \\ 0, & \text{otherwise}. \end{cases}$$

(16)

In the case of $\bar{A}_m \neq \emptyset$, the content outage probability for user $u_m$ in $P_1$ also depends on the applied joint coding scheme. Therefore, after generating the optimal solution of $P_2$ as Eq. (16), the optimal joint coding scheme will then be applied on $\mathcal{A}_m^*$ for user $u_m$. An example of $\mathcal{M}_{b}^{fm}$ is shown in Fig. 2. It can be seen that $u_1$, $u_3$ and $u_4$ are $\mathcal{M}_{b}^{fm}$-saturated, and thus are not in outage. The outage state of $u_2$, however, is determined by the sum capacity achieved by the applied coding scheme.

C. Distributed Fog-AP Selection Algorithm

Based on the fast message passing method in [19], a distributed Fog-AP selection algorithm is proposed to solve problem $P_2$ in this subsection. As discussed in Section III-B, the optimal Fog-AP selection can be obtained by finding the fairness maximum $b$-matching $\mathcal{M}_{b}^{fm}$. By representing the maximum $b$-matching problem in a factor graph, it can be mapped into a marginal probability computation problem on the probabilistic graphical model. Therefore, the maximum $b$-matching problem can then be solved by passing messages between the adjacent vertices on the probabilistic graphical model. The fairness is guaranteed by adding randomness in the message passing process. The basic message passing algorithm for solving the fairness maximum $b$-matching problem is similar to the algorithm in [20], whose total computation cost scales as $O(NM^2)$.

To reduce the memory usage to $O(M + N)$, an improved message passing algorithm proposed in [19] is applied to solve the fairness maximum $b$-matching problem. A key step of this algorithm is the selection operation, which finds the $k$-th largest element in a set $\mathcal{X}$ for a given $k$. The selection operation over $\mathcal{X}$ can be written as

$$\sigma_k(\mathcal{X}) = \{x \in \mathcal{X} : |\{t \in \mathcal{X} : t \geq x\}| = k\}.$$ 

(17)

For convenience, we define

$$\tilde{Q} = \begin{bmatrix} O & Q \\ Q^T & O \end{bmatrix},$$

(18)
where $Q$ is the 1 bit quantized CSI matrix, and

$$
\hat{X} = \begin{bmatrix}
O & X \\
X^T & O
\end{bmatrix},
$$

(19)

where $X$ is the decision matrix in problem P2. The message passing algorithm maintains a belief value for each edge in $G(V \cup A, E)$, which can be denoted by a matrix $B$. Similarly, we define

$$
\hat{B} = \begin{bmatrix}
O & B \\
B^T & O
\end{bmatrix}.
$$

(20)

The belief value for the edge $e_{ij} = (v_i, v_j)$ at iteration $t$ is denoted by the entry $[\hat{B}]_{ij}^t$, where $v_i = u_m$ for $i = 1, 2, \ldots, M$ and $v_i = a_n$ for $i = M + 1, M + 2, \ldots, M + N$. $[\hat{B}]_{ij}^t$ is updated with the following rule

$$
[\hat{B}]_{ij}^t = [Q]_{ij} - \sigma_L \left( \left\{ [\hat{B}]_{jk}^{t-1} : v_k \in U \cup A, k \neq i \right\} \right).
$$

(21)

Define $\mu_j$ as the negation of the $b(v_j)$-th selection and $\nu_j$ for the $(b(v_j) + 1)$-th selection. $\mu_j$ and $\nu_j$ are updated with the following rules

$$
\begin{align*}
\mu_j^t &= -\sigma_{b(v_j)} \left( \left\{ [\hat{B}]_{jk}^{t-1} : v_k \in U \cup A \right\} \right) ; \\
\nu_j^t &= -\sigma_{b(v_j)+1} \left( \left\{ [\hat{B}]_{jk}^{t-1} : v_k \in U \cup A \right\} \right).
\end{align*}
$$

(22)

Let $\hat{x}_{ij}^t$ denote the entry at the $i$-th row and $j$-th column of $\hat{X}$ at the $t$-th iteration. The resulting belief searching rule can then be written as

$$
[\hat{B}]_{ij}^t = \begin{cases}
[Q]_{ij} + \mu_j^t, & \hat{x}_{ji}^t = 1; \\
[Q]_{ij} + \nu_j^t, & \text{otherwise}.
\end{cases}
$$

(23)

The estimation of $\hat{X}$ after each iteration will be updated as

$$
\hat{x}_{mn}^t = \begin{cases}
1, & \mu_i^t \leq [\hat{B}]_{ij}^{t-1}; \\
0, & \text{otherwise}.
\end{cases}
$$

(24)

It is shown in [19] that this algorithm converges if there exists a valid $b$-matching solution. The details of this algorithm are summarized in Algorithm [1].

To further reduce the running time, a variation of Algorithm [1] is introduced, where Steps 7-8 are improved. Since $\hat{Q}$ does not change during the message passing, the algorithm can compute the index cache $I \in \mathbb{N}^{(M+N) \times c}$ with cache size $c$ initially, where $[I]_{ik}$ is the index of the $k$-th largest weight connected to node $v_i$.

For $\kappa = [I]_{ik}$,

$$
[\hat{Q}]_{in} = \sigma_k \left( \left\{ [Q]_{ij} : v_j \in U \cup A \right\} \right).
$$

(25)
Algorithm 1 Distributed Fog-AP Selection

1: $\mu_0^j, \nu_0^j \leftarrow 0$ for $j = 1, 2, \ldots, M + N$
2: $\tilde{X}^0 \leftarrow \emptyset$
3: $t \leftarrow 1$
4: while not converge do
5:  for $j = 1, \ldots, M + N$ do
6:   $\tilde{x}^t_{jk} \leftarrow 0$, for $k = 1, 2, \ldots, M + N$
7:   $\mu_j^t \leftarrow -\sigma_b(j)\left(\{[\tilde{B}]_{jk}^{t-1} : v_k \in \mathcal{U} \cup \mathcal{A}\}\right)$
8:   $\nu_j^t \leftarrow -\sigma_b(j)+1\left(\{[\tilde{B}]_{jk}^{t-1} : v_k \in \mathcal{U} \cup \mathcal{A}\}\right)$
9:   for all $\{k : \mu_j^t \leq [\tilde{B}]_{jk}^{t-1}\}$ do
10:    $\tilde{x}^t_{jk} \leftarrow 1$
11:  end for
12: end for
13: Delete $\tilde{X}^{t-1}$, $\mu_j^{t-1}$, and $\nu_j^{t-1}$
14: $t \leftarrow t + 1$
15: end while
16: Output $\tilde{X}$

As for $\theta = d_k$ and entry $\nu_j^t = \sigma_k(\nu_j^t : v_j \in \mathcal{U} \cup \mathcal{A})$, the $\nu^t$ values are similarly sorted and stored in the index vector $d \in \mathbb{N}^{M+N}$ after each iteration. After updating process of each step, we maintain a set $\mathcal{B}$ of the greatest $b(v_j) + 1$ beliefs obtained so far. Thus, the current estimation for $\mu_j^t$ and $\nu_j^t$ at each stage are

\[
\begin{aligned}
\hat{\mu}_j^t &= \sigma_{b(v_j)}(\mathcal{B}), \\
\hat{\nu}_j^t &= \min(\mathcal{B}).
\end{aligned}
\]  

(26)

At each iteration, the greatest possible unseen belief is bounded as the sum of the least weight seen so far from the sorted weight cache and the least $\nu$ value so far from the $\nu$ cache. The sufficient selection procedure will exit when $\hat{\nu}_j^t$ becomes less than or equals to the sum, since further comparisons are unnecessary. The details of this algorithm is shown in Algorithm 2.

Remark 1: Algorithm 1 is the key result of the proposed $b$-matching approach, which will be implemented in each Fog-AP. The procedure in each coherence time is listed as follows:

1) At the beginning of each coherence time, the user will feedback one bit CSI according to the channel estimation result to show if the Fog-AP is in outage for a user.
Algorithm 2 Sufficient Selection
1: \( k \leftarrow 1 \)
2: \( \rho \leftarrow \infty \)
3: \( \mathcal{B} \leftarrow \emptyset \)
4: \( \hat{\mu}_t^j \leftarrow -\infty \)
5: \( \hat{\nu}_t^j \leftarrow -\infty \)
6: while \( \hat{\nu}_t^j < \rho \) do
7: \hspace{1em} if \( k \leq c \) then
8: \hspace{2em} \( \kappa \leftarrow [I]_{jk} \)
9: \hspace{2em} if \( \kappa \) is unvisited and \( \tilde{\mathcal{B}}_{j\kappa}^{t-1} > \min(\mathcal{B}) \) then
10: \hspace{3em} \( \mathcal{B} \leftarrow (\mathcal{B} \setminus \min(\mathcal{B})) \cup [\tilde{\mathcal{B}}_{j\kappa}^{t-1}] \)
11: \hspace{2em} end if
12: \hspace{1em} end if
13: \hspace{1em} \( \theta \leftarrow d_k \)
14: \hspace{1em} if \( \theta \) is unvisited and \( \tilde{\mathcal{B}}_{j\theta}^{t-1} > \min(\mathcal{B}) \) then
15: \hspace{2em} \( \mathcal{B} \leftarrow (\mathcal{B} \setminus \min(\mathcal{B})) \cup [\tilde{\mathcal{B}}_{j\theta}^{t-1}] \)
16: \hspace{1em} end if
17: \hspace{1em} \( \rho \leftarrow \tilde{Q}_{j\kappa} + \nu_{\theta}^{t-1} \)
18: \hspace{1em} \( \tilde{\mu}_t^j \leftarrow \sigma_{b(v_j)}(\mathcal{B}) \)
19: \hspace{1em} \( \tilde{\nu}_t^j \leftarrow \sigma_{b(v_j)+1}(\mathcal{B}) \)
20: \hspace{1em} \( k \leftarrow k + 1 \)
21: end while
22: \( \mu_t^j \leftarrow \tilde{\mu}_t^j \)
23: \( \nu_t^j \leftarrow \tilde{\nu}_t^j \)

2) Algorithm [1] will be executed between Fog-APs and users to select the Fog-APs for each user. This algorithm will converge very quickly because of its low complexity.

3) After Algorithm [1] the data will be transmitted from the selected Fog-APs to each user.

IV. CONTENT OUTAGE PERFORMANCE ANALYSIS AND DMR OPTIMAL MSR CODE

In this section, the outage probability will be analyzed for problem P2. With the closed-form formula for content outage probability, the optimal \( K_m \) can be obtained, so that the optimization problem P1 is solved.
A. Conditional Content Outage Probability

As discussed before, $\mathcal{M}_b^u$ may not always exist, so that $k_m$ may be smaller than $K_m$ for user $u_m \in \mathcal{U}$. Thus, we need first calculate the tight upper bound of the content outage probability under the condition that $k_m$ non-outage Fog-APs can be accessed by user $u_m$.

With a generated $\mathcal{M}_b^{\text{fm}}$, the conditional content outage probability of user $u_m$ is written by

$$
p_{\text{con}}^m(R_m|D_{K_m,k_m}) = \Pr\left\{ \sum_{a_n \in A_m} I_{mn} < R_m \bigg| D_{K_m,k_m} \right\},
$$

(27)

where

$$
D_{K_m,k_m} = \{C_{m1}, \ldots, C_{mK_m} | C_{m1} \geq \alpha^*_m, \ldots, C_{mk_m} \geq \alpha^*_m, \\
C_{m,k_m+1} < \alpha^*_m, \ldots, C_{mK_m} < \alpha^*_m \}.
$$

(28)

In the following, the saddle-point approximation in [21], [22] is applied to derive a tight upper bound of Eq. (27). Let $\rho_m = 1 - \frac{k_m}{K_m}$, and define a symbol “≲” as

$$
f(x) \lesssim g(x) \iff \begin{cases} f(x) \leq g(x); \\
\lim_{x \to \infty} \frac{f(x)}{g(x)} = 1.
\end{cases}
$$

(29)

We then have the following theorem.

**Lemma 1:** The exponentially tight upper bound of $p_{\text{con}}^m(R_m|D_{K_m,k_m})$ is given by

$$
p_{\text{con}}^m(R_m|D_{K_m,k_m}) \lesssim p_{\text{upper}}^m(R_m|D_{K_m,k_m})
$$

(30)

where

$$
\psi = \frac{1}{\sqrt{2\pi K_m \sigma^2 \lambda^*}},
$$

(31)

$$
\begin{align*}
J_1(\gamma) &= \ln p_m^{\text{fm}} q_m^{1-\rho_m}; \\
J_0(\gamma) &= \frac{1}{\gamma} + (1 - \rho_m) \ln \Gamma\left(1 - \lambda^*, \frac{e\alpha^*_m}{\gamma}\right) \\
&\quad + \rho_m \ln \left(\Gamma\left(1 - \lambda^*, \frac{1}{\gamma}\right) - \Gamma\left(1 - \lambda^*, \frac{e\alpha^*_m}{\gamma}\right)\right).
\end{align*}
$$

(32)

In Eqs. (31) (32), $\lambda^*$ and $\sigma^2$ satisfy Eq. (35) and Eq. (36), respectively. In these equations,

$$
\Gamma(z, a) = \int_a^{\infty} e^{-t} t^{z-1} dt
$$

(33)

is the incomplete Gamma function, and

$$
G_{m,n}^{p,q} \left( z \bigg| \begin{array}{c}
a_1, \ldots, a_p \\
b_1, \ldots, b_q
\end{array} \right) = \frac{1}{2\pi i} \oint_{L} \prod_{j=1}^{m} \Gamma(b_j - s) \prod_{j=1}^{n} \Gamma(1 - a_j + s) \prod_{j=m+1}^{p} \Gamma(a_j - s) \prod_{j=n+1}^{q} \Gamma(1 - b_j + s) z^s ds,
$$

(34)

is the Meijer’s $G$-function [23].

**Proof:** See Appendix B. 

\[ \square \]
(35)

\[
\alpha_m^* = \ln \gamma + (1 - \rho_m) \ln \frac{e^{\alpha_m^*}}{\gamma} + (1 - \rho_m) \frac{G^{2,0}_{2,3} \left( \frac{e^{\alpha_m^*}}{\gamma} \right)}{\Gamma \left( 1 - \lambda^* \frac{e^{\alpha_m^*}}{\gamma} \right)} \frac{\left| \begin{array}{c} 1, 1, - \\ 0, 0, 1 - \lambda^* \end{array} \right|}{\Gamma \left( 1 - \lambda^* \frac{e^{\alpha_m^*}}{\gamma} \right)}
\]

\[
\rho_m \left[ \frac{\Gamma \left( 1 - \lambda^* \frac{e^{\alpha_m^*}}{\gamma} \right) \ln \frac{1}{\gamma} - \Gamma \left( 1 - \lambda^* \frac{e^{\alpha_m^*}}{\gamma} \right) \ln e^{\alpha_m^*} + G^{2,0}_{2,3} \left( \frac{e^{\alpha_m^*}}{\gamma} \right)}{\Gamma \left( 1 - \lambda^* \frac{e^{\alpha_m^*}}{\gamma} \right)} \right]
\]

\[
\sigma^2 = \rho_m \left[ \frac{\Gamma \left( 1 - \lambda^* \frac{e^{\alpha_m^*}}{\gamma} \right) \ln \frac{1}{\gamma} - \Gamma \left( 1 - \lambda^* \frac{e^{\alpha_m^*}}{\gamma} \right) \ln e^{\alpha_m^*}}{\Gamma \left( 1 - \lambda^* \frac{e^{\alpha_m^*}}{\gamma} \right)} \right]
\]

(36)

\[
\sigma^2 = \rho_m \left[ \frac{\Gamma \left( 1 - \lambda^* \frac{e^{\alpha_m^*}}{\gamma} \right) \ln \frac{1}{\gamma} - \Gamma \left( 1 - \lambda^* \frac{e^{\alpha_m^*}}{\gamma} \right) \ln e^{\alpha_m^*}}{\Gamma \left( 1 - \lambda^* \frac{e^{\alpha_m^*}}{\gamma} \right)} \right]
\]

\[
- \rho_m \left[ \frac{\Gamma \left( 1 - \lambda^* \frac{e^{\alpha_m^*}}{\gamma} \right) \ln \frac{1}{\gamma} - \Gamma \left( 1 - \lambda^* \frac{e^{\alpha_m^*}}{\gamma} \right) \ln e^{\alpha_m^*}}{\Gamma \left( 1 - \lambda^* \frac{e^{\alpha_m^*}}{\gamma} \right)} \right]
\]

\[
+ \frac{G^{2,0}_{2,3} \left( \frac{e^{\alpha_m^*}}{\gamma} \right)}{\Gamma \left( 1 - \lambda^* \frac{e^{\alpha_m^*}}{\gamma} \right)} \frac{\left| \begin{array}{c} 1, 1, - \\ 0, 0, 1 - \lambda^* \end{array} \right|}{\Gamma \left( 1 - \lambda^* \frac{e^{\alpha_m^*}}{\gamma} \right)} \left( \frac{G^{2,0}_{2,3} \left( \frac{e^{\alpha_m^*}}{\gamma} \right)}{\Gamma \left( 1 - \lambda^* \frac{e^{\alpha_m^*}}{\gamma} \right)} \right)^2
\]

\[
\] 

B. Content Outage Probability

The content outage probability of the proposed RBG based \( b \)-matching approach is analyzed in this subsection. The exact performance analysis is intractable, so that we focus on the first order approximations in both high and low SNR regimes.

To analyze the content outage probability, we need to study the properties of the maximum \( b \)-matching on RBG first. Due to space limitation, we introduce the following lemma, where the detailed proof can be found in [24].
Lemma 2:

\[ |M_m| = \min_{X \subseteq V} \{ b(V \setminus X) + |E(X)| \}. \] (37)

Let \( \{K_{m:M}\}_{m=1}^{M} \) be the ordered sequence of \( \{K_m\}_{m=1}^{M} \), where \( K_{1:M} \leq \cdots \leq K_{M:M} \). Denote the ordered sequence of \( \{u_m\}_{m=1}^{M} \) as \( \{u_{m:M}\}_{m=1}^{M} \), where \( u_{m:M} \) has the same order as \( K_{m:M} \). Define

\[ K_{\text{sum}} = \sum_{m=1}^{M} K_m. \] (38)

The following lemma can then be established from Lemma 2.

Lemma 3: Let \( G((U \cup A, E) \) be a bipartite graph with \( |U| = M \) and \( |A| = N \). \( b(v) \) is defined as Eq. (12). If \( M_m \) exists such that:

1) the only unsaturated vertex is in \( \{u_{i:M}\}_{i=m}^{M} \), the cardinality of \( E \) is given by

\[ \Phi_1 = (M - 1)N + K_{m:M} - 1; \] (39)

2) \( U \) has only one unsaturated vertex, the cardinality of \( E \) is given by

\[ \Phi_2 = (M - L) \left( \left\lceil \frac{K_{\text{sum}}}{L} \right\rceil - 1 \right) + K_{\text{sum}} - 1. \] (40)

Proof: See Appendix C.

Based on Lemma 1 and Lemma 3, the content outage probability is summarized in the following theorem.

Theorem 1: If \( N > \frac{\Phi_2 - \eta K_m}{M-1} \), the content outage probability of \( u_{m:M} \) in high SNRs is given by

\[ p_{\text{out}}^{\text{M}}(R_{m:M}) = \sum_{\kappa=\Phi_2}^{\Phi_1} \left( \begin{array}{c} N \\ \kappa \end{array} \right) p_{m:M}^{\text{con}}(R_{m:M} | D_{K_{m:M},N-\kappa}) p_{m:M}^{\kappa} + O\left( p_{m:M}^{N - \kappa} \right). \] (41)

If \( N < \frac{\Phi_2 - \eta K_m}{M-1} \), the content outage probability of \( u_{m} \) in high SNRs is given by

\[ p_{\text{out}}^{\text{m}}(R_{m}) = \left( \frac{MN}{\Phi_2} \right) p_{m}^{\text{con}}(R_{m} | D_{K_{m},\eta K_m}) p_{m}^{MN - \Phi_2} + O\left( p_{m}^{N - \Phi_2 + \eta K_m} \right). \] (42)

If \( N = \frac{\Phi_2 - \eta K_m}{M-1} \), the content outage probability of \( u_{m:M} \) in high SNRs is given by the summation of Eq. (41) and Eq. (42).

In low SNRs, the content outage probability of \( u_{m} \) is then given by

\[ p_{\text{out}}^{\text{m}}(R_{m}) = p_{m}^{N} + N p_{m}^{\text{con}}(R_{m} | D_{K_{m},1}) p_{m}^{N-1} q_{m} + O(q_{m}). \] (43)

Proof: See Appendix D.
Fig. 3. The OER $\mathcal{R}_{OE}^{OE}(R_1, R_2; \gamma)$ for a coded caching scheme in Fog-RAN.

Fig. 4. The OER and DMR for coded caching scheme in the Fog-RAN with $M = 10$, $N = 5$, $L = 4$, and $r_m = (0.9, 0.9)$.

C. DMR Optimal MSR Code

To illustrate the solution of P1 in Eq. (9), we define the outage exponent region (OER), which specifies the set of content outage exponents that are simultaneously achievable by the users for the specific coded caching scheme. As a matter of fact, the outage exponent proposed in [18] is defined as

$$E_m(R_m, \gamma) = -\frac{\partial \ln p_{out}^{out}(R_m)}{\partial \ln \gamma}. \quad (44)$$

Similar to error exponent region in [25], the OER can be formally defined in the following.

**Definition 1:** For a coded caching scheme with $\mathbf{R} = (R_m)_{m=1}^{M}$, the OER, denoted by $\mathcal{R}_{OE}(\mathbf{R}; \gamma)$, consists of all vectors of outage exponents $(E_m(R_m, \gamma))_{m=1}^{M}$, which can be achieved by at least one feasible solution of Eq. (11).

Fig. 3 shows an example of the OER for a coded caching scheme with $(R_1, R_2)$. In contrast to the unique capacity region, one OER corresponds to one coded caching scheme. Based
Theorem 1: For a coded caching scheme with $R$, the best achievable bound of the $OER$ $R_{OER}(\gamma)$ is given by the vector $(E_m(R_m, \gamma))_{m=1}^M$, which can be obtained by plugging Eq. (41) and Eq. (42) into Eq. (44).

The DMT for a user $u_m$ is closely related to the outage exponent as follows [18]:

$$
\lim_{\gamma \to \infty} E_m(R_m, \gamma) = d_m(r_m),
$$

where $d_m(r_m)$ is the diversity gain for the multiplexing gain $r_m$. Recalling Eq. (45), we have the following corollary.

**Corollary 1:** The conditional DMT for $u_m$ is given by

$$
d_m^{con}(r_m | D_{K_m}) = K_m (1 - \rho_m) \left(1 - \frac{r_m}{K_m}\right).
$$

Similar to OER, the DMT can be generalized to Fog-RANs, i.e., the diversity-multiplexing region (DMR), which is defined in the following.

**Definition 2:** For a coded caching scheme with $r = (r_m)_{m=1}^M$, the DMR, denoted by $R_{DM}(r)$, consists of all vectors of diversity gains $(d_m(r_m))_{m=1}^M$, which can be achieved by at least one feasible solution of Eq. (11).

As shown in Fig. [3], the DMR is the asymptotic case of the OER when $\gamma$ tends to infinity. According to Definition [2] and Corollary [1], the achieved DMR is obtained from Theorem [1].

**Theorem 3:** For a coded caching scheme with $r = (r_m)_{m=1}^M$, the best achievable bound of the DMR $R_{DM}(r)$ is given by the vector $(d_m^*(r_m))_{m=1}^M$ for $0 < r_m \leq K_m$ which satisfies

$$
d_m^*(r_m) = \begin{cases} 
N \left(1 - \frac{r_m}{K_m}\right), & \text{if } N \geq \frac{\Phi_2 - \eta K_m}{M-1}; \\
(MN - \Phi_2 + \eta K_m) \left(1 - \frac{r_m}{K_m}\right), & \text{if } N < \frac{\Phi_2 - \eta K_m}{M-1}.
\end{cases}
$$

According to [18], the optimal DMT for point-to-point parallel fading channels is given by

$$
d^*(r) = N \left(1 - \frac{r}{N}\right),
$$
where $N$ is the number of subchannels. In this context, the optimal $K_m$, denoted by $K^*_m$, in P1 can be determined by

$$K^*_m = \frac{R_m}{\sum_{i=1}^{M} R_i} N.$$  \hfill (50)

if the RBG based $b$-matching scheme is required to achieve the optimal DMR. Therefore, in the optimal DMR, each user shares the total multiplexing gain according to $R$, while achieving the full frequency diversity. The DMR and the DMT curve of the user $u_m$ are respectively shown in Fig. 4 and Fig. 5 where the OER approaches the DMR as SNR tends to infinity.

As discussed in Section II-B, the performance of MSR code is determined by $(\alpha^*_m, \beta^*_m)$. In this context, the requirement of the MSR code can be obtained if it needs to achieve the optimal DMR performance. According to (50), the user $u_m$ will access $K^*_m = \frac{R_m}{\sum_{i=1}^{M} R_i} N$ Fog-APs to download the interested content $s_m$. Furthermore, recalling Eq. (2), we have

$$\alpha^*_m = \frac{R_m}{K_m} = \frac{\sum_{i=1}^{M} R_i}{N}$$  \hfill (51)

for $m = 1, \ldots, M$. It can be seen that with the optimal $K^*_m$, $\alpha^*_m$, the information transmitted from a Fog-AP to user $u_m$, is the same as each other for the required content $s_m$ in $S$. In Eq. (2), $\beta^*_m$ is given by

$$\beta^*_m = \frac{D_m}{D_m - K^*_m + 1} \frac{\sum_{i=1}^{M} R_i}{N}.$$  \hfill (52)

Thus, the optimal $D_m$, denoted by $D^*_m$, is lower bounded by

$$D^*_m \geq K^*_m = \frac{\sum_{i=1}^{M} R_i}{N}.$$  \hfill (53)
It can be seen that it requires at least $D^*_m = \frac{R_m}{\sum_{i=1}^{M} R_i} N$ Fog-APs to repair the content $s_m$ in a new Fog-AP. If we choose $D^*_m$ as the lower bound, the transmitted information is given by

$$\beta^*_m = D^* \frac{\sum_{i=1}^{M} R_i}{N} = R_m.$$  \hspace{1cm} (54)

In this context, we have the following theorem.

**Theorem 4:** To achieve the optimal DMR performance of the coded caching scheme in Fog-RAN, the parameters of MSR code for $u_m$ with the content $s_m$ can be given by

$$(N, K^*_m, D^*_m) = \left( N, \frac{R_m}{\sum_{i=1}^{M} R_i} N, \frac{R_m}{\sum_{i=1}^{M} R_i} N \right).$$ \hspace{1cm} (55)

The code performance is characterized by

$$(\alpha^*_m, \beta^*_m) = \left( \frac{\sum_{i=1}^{M} R_i}{N}, R_m \right).$$ \hspace{1cm} (56)

The MSR code defined in Theorem 4 achieves the Pareto optimal DMR with fairness constraints. Specifically, the fairness is guaranteed because the maximum achievable multiplexing gain for each user is proportional to the size of content stored in Fog-APs, i.e., $K_m \propto R_m$ for $m = 1, \ldots, M$. According to Theorem 3 for any given multiplexing gain $r_m \in (0, K_m]$, the $b$-matching approach achieves the maximum diversity gain. Moreover, any change of the MSR code parameters for all the users in $U$ cannot improve the performance of one user without hurting another user, which fulfills the definition of Pareto optimal [26].

V. Simulation Results

In this section, some simulation examples are presented. In the first group of simulations, the upper bound of the conditional outage probability is verified, which lays the foundation of the proposed framework. The second group of simulations verifies the outage performance. The last group of simulation compares the performance with different schemes and system parameters. In these figures, the DMT curves are plotted to illustrate that they are in parallel with the content outage probability curves (i.e., they have the same slope) in the high SNR regime.

For the first group of simulation results, a large number of samples are generated at each SNR value. The number of outage events can be obtained under the condition that the transmitter knows 1 bit CSI by computing the instantaneous channel capacity. The conditional outage probability $p_m^{\text{cop}} (R_m | D_{K_m k_m})$ is then calculated. The theoretical approximations are calculated by applying the results in Lemma 1. Fig. 2 compares the simulation results and the theoretical curves of the conditional outage probability. It can be seen that the proposed
tight upper bound is nearly identical with the simulation results in high SNR regime. For comparison, the curve of non-conditional outage probability is also plotted. Fig. 7 presents the curves of conditional outage exponent and conditional DMT at a given $R_m = 2$. Clearly, the conditional outage exponent is an increasing function of SNR for a fixed $R_m$. The proposed theoretical curve is approaching the simulation one when SNR increases. In Fig. 7 the conditional DMT is plotted as a constant which is much larger than the outage exponent at low SNRs. However, the conditional outage exponent approaches the conditional DMT as SNR tends to infinity. Therefore, the proposed conditional outage exponent can be used to estimate the decreasing slope of conditional outage probabilities with 1 bit CSI.
For the second group, a large number of samples are generated according to the $b$-matching approach of the Fog-AP selection problem. In the proposed allocation scheme, the approximation of the outage probabilities are calculated by the formulas in Theorem 1. The asymptotic line $\text{SNR}^{d_m}(r_m)$ is also plotted to compare the slope of these curves in high SNRs. Fig. 8 compares the simulation results and the theoretical curves of the proposed $b$-matching approach. These simulation examples consider the Fog-AP selection problem with $M = 10$ users and $N = 5$ Fog-APs, where each Fog-AP can only support $L = 4$ users at most and each user will access $K_m = 2$ Fog-APs for fulfilling the requirement of MSR codes. In addition, the theoretical results of the content outage probability with Eq. (41) and
Eq. (43) are also illustrated simultaneously. The fixed target rate schemes, i.e., \( R_m = 2 \), are plotted in Fig. 8. Since the conditional outage probability \( p_{\text{cop}}(R_m | D_{K_m k_m}) \) is calculated by utilizing the results in Lemma 1, the theoretical calculations are still suitable for only high SNR regime. It can be seen that the theoretical approximation in both Eq. (41) and Eq. (43) approach the simulation results when SNR increases. Moreover, for a coded caching scheme with \( r = (r_m)_{m=1}^M = 0 \), the best achievable bound of the DMR \( R^{\text{DM}}(r) \) given by the vector in Eq. (47) satisfying \( d^*_m(r_m) = 5 \) is also demonstrated in Fig. 8. In Fig. 9 and Fig. 10, the dynamic rate scenario is considered, where the multiplexing gain \( r_m \) is set to 0.9 and 0.6, respectively. The theoretical approximation is computed by Eq. (41) and Eq. (43) in high and low SNR regimes, from \(-10 \text{ dB}\) to \(25 \text{ dB}\) under \( r_m = 0.9 \) and from \(-10 \text{ dB}\) to \(15 \text{ dB}\) under \( r_m = 0.6 \). The curves indicate that the theoretical approximation is nearly identical with the numerical simulation results, particularly for Eq. (43) in low SNR range. The best achievable bounds of the DMR \( R^{\text{DM}}(r) \) calculated by Eq. (47) are given as \( d^*_m(r_m) = 2.75 \) and \( d^*_m(r_m) = 3.5 \), respectively.

The last group of simulation compares the performance with different schemes and system parameters. According to [27], the commonly used coded caching schemes are summarized in Table I. As the parameters of MDS code is quite different from our schemes. The results in Fig. 11 show the performance comparison of MBR codes and MSR codes with different parameters. It can be seen that the outage probability of the MSR code outperforms the MBR code with the same coding parameters. Moreover, many parameters could influence the performance of the \( b \)-matching scheme. Hence, Fig. 12 shows the simulation results.
with different parameters. It can be seen that different parameters yield different outage performance. According to the theoretical analysis, the slope of the outage probability is determined in Theorem 3. Therefore, the outage probability with $N = 4, L = 3, K_m = 2, M = 6$ is almost the same as that with $N = 4, L = 4, K_m = 2, M = 8$. However, the outage probability with $N = 5, L = 4, K_m = 2, M = 10$ is different from that with $N = 5, L = 6, K_m = 3, M = 10$ as $K_m$ is different in these curves.

### Table I

| Scheme | $n$ | $k$ | $d$ | $\alpha_m$ | $\beta_m$ |
|--------|-----|-----|-----|------------|------------|
| MDS    | $N$ | $K_m$ | $-$ | $\frac{R_m}{K_m}$ | $-$ |
| MBR    | $N$ | $K_m$ | $D_m$ | $\frac{2D_m - R_m + 1}{K_m}$ | $\frac{2D_m}{K_m} + \frac{R_m}{K_m}$ |
| MSR    | $N$ | $K_m$ | $D_m$ | $\frac{R_m}{K_m}$ | $\frac{D_m}{K_m - K_m + 1} + \frac{R_m}{K_m}$ |

### VI. Conclusions

This paper proposed a $b$-matching approach for coded caching in Fog-RANs, which combines the advantages of coded caching scheme and RBG based fairness maximum $b$-matching. The combinatorial structure of Fog-AP selection is explored, and the optimal parameters of MSR code is obtained. The fast message passing algorithm is proposed to solve the optimal Fog-AP selection problem, which reduce the memory usage to $O(M + N)$, compared to conventional message passing method. It was shown by theoretical derivations that all of the
Fig. 12. The outage probability of the $b$-matching approach for different system parameters.

users fairly share the total multiplexing gain while achieving the full frequency diversity, i.e., the proposed scheme is optimal overall caching and transmission schemes from the DMR perspective. The simulation results illustrate the accuracy of the theoretical derivations, and verify that the optimal outage performance of the proposed scheme. Therefore, the $b$-matching approach provides an elegant theoretical framework for designing coded caching transmission scheme in Fog-RANs.

APPENDIX A

PRELIMINARIES ON RBG AND $b$-MATCHING

In this appendix, we introduce some fundamental concepts of RBG [28]. A bipartite graph $\mathcal{G}(\mathcal{V}_1 \cup \mathcal{V}_2, \mathcal{E})$ is a graph on which the set of vertices $\mathcal{V}$ admits a partition into two classes $\mathcal{V}_1$ and $\mathcal{V}_2$ such that every edge in the edge set $\mathcal{E}$ has its ends in different classes. Specifically, for any edge $e = (v, v')$ in $\mathcal{E}$, we have $v \in \mathcal{V}_1$ and $v' \in \mathcal{V}_2$. For convenience, denote $v \in e$ if $e = (v, v')$. A bipartite graph with $M$ and $N$ vertices in different partition classes is called complete, denoted by $K_{MN}$, if every two vertices from different partition classes are adjacent. Similar to [28], the probability space of a RBG is defined by $\mathcal{G}(K_{MN}; P)$, which means that whether the edge of $K_{MN}$ appears follows the probability measure $P$. A vertex $v'$ is called a neighbor of vertex $v$, if they are adjacent. All of the neighboring vertices of a given vertex $v$ are denoted by the set $\mathcal{N}(v) \subseteq \mathcal{V}$. The cardinality of $\mathcal{N}(v)$ is referred to as the degree of
the vertex $v$, denoted by $\deg_G(v)$. The adjacent vertex set of $\mathcal{X} \subseteq \mathcal{V}$ is then defined by

$$\mathcal{N}(\mathcal{X}) = \bigcup_{v \in \mathcal{X}} \mathcal{N}(v).$$

(57)

For convenience, we also denote

$$\mathcal{E}(\mathcal{X}) = \bigcup_{v \in \mathcal{X}} \mathcal{E}(v),$$

(58)

and

$$b(\mathcal{X}) = \bigcup_{v \in \mathcal{X}} b(v)$$

(59)

for any $\mathcal{X} \subseteq \mathcal{V}$.

Based on the bipartite graph, the $b$-matching can be defined as follows [24]. Let $b(v)$ be a non-negative integer associated with the vertex $v \in \mathcal{V}$, and $\mathcal{E}(v)$ be the set of edges incident with $v$. The $b$-matching $\mathcal{M}_b$ is the subset of $\mathcal{E}$ satisfying

$$|\mathcal{M}_b \cap \mathcal{E}(v)| \leq b(v).$$

(60)

If $|\mathcal{M}_b \cap \mathcal{E}(v)| = b(v)$, the vertex $v$ is referred to as the $\mathcal{M}_b$-saturated. The maximum $b$-matching, denoted by $\mathcal{M}_m^b$, is the $b$-matching with the maximum number of edges. Especially, the $\mathcal{V}_1$-perfect $b$-matching, denoted by $\mathcal{M}_p^b$, means that the equality of Eq. (60) holds for every $v \in \mathcal{V}_1$.

**APPENDIX B**

**PROOF OF LEMMA**

Define a sequence of random variables $\{Z_{mn}\}_{n=1}^{K_m}$. The cumulative distribution function of $Z_{mn}$ for $1 \leq n \leq k_m$ is given by

$$F_{Z_{mn}}(z) = \Pr\{I_{mn} < z| \mathcal{D}_{K_m k_m}\} = \Pr\{I_{mn} < z| I_{mn} \geq \alpha_m^*\}$$

$$= \begin{cases} 1 - \frac{1}{q_m} \exp \left(-\frac{e^z - 1}{\gamma}\right), & z \geq \alpha_m^*; \\ 0, & \text{otherwise}. \end{cases}$$

(61)

For $k_m + 1 \leq n \leq K_m$, the cumulative distribution function of $Z_{mn}$ is given by

$$F_{Z_{mn}}(z) = \Pr\{I_{mn} < z| \mathcal{D}_{K_m k_m}\} = \Pr\{I_{mn} < z| I_{mn} < \alpha_m^*\}$$

$$= \begin{cases} \frac{1}{p_m} \left[1 - \exp \left(-\frac{e^z - 1}{\gamma}\right)\right], & z < \alpha_m^*; \\ 1, & \text{otherwise}. \end{cases}$$

(62)

Let $X_n = \alpha_m^* - Z_{mn}$ and $Y_{K_m} = \sum_{n=1}^{K_m} X_n$, the conditional outage probability in Eq. (27) can be rewritten as $p_m^{\text{con}}(R_m| \mathcal{D}_{K_m k_m}) = \Pr\{Y_{K_m} > 0\}$. 
Recalling that the elements of \( \{ h_{mn} \}_{n=1}^{K_m} \) are independent with the identical distribution of \( CN(0,1) \). Let \( 1 \leq n_1 \leq k_m, \ k_m + 1 \leq n_2 \leq K_m \), the cumulant-generating function of \( Y_{K_m} \) is then given by

\[
\Lambda(\lambda) = \lim_{K_m \to \infty} \frac{1}{K_m} \ln \mathbb{E} \{ e^{\lambda Y_{K_m}} \} = \alpha_m^* \lambda + \ln \left( \mathbb{E} \{ e^{-\lambda Z_{mn2}} \} \right)^{1-\rho_m} + \ln \left( \mathbb{E} \{ e^{-\lambda Z_{mn1}} \} \right)^{1-\rho_m}
\]

\[
= (\alpha_m^* - \ln \gamma) \lambda + \frac{1}{\gamma} - \ln p_m(\lambda, y \lambda) + (1 - \rho_m) \ln \Gamma \left( 1 - \lambda, \frac{e^{\alpha_m^*}}{\gamma} \right) + \rho_m \ln \left( \Gamma \left( 1 - \lambda, \frac{1}{\gamma} \right) - \Gamma \left( 1 - \lambda, \frac{e^{\alpha_m^*}}{\gamma} \right) \right)
\]

(63)

where \( \rho_m = 1 - \frac{k_m}{K_m} \). Considering the relation between the cumulant-generating function and the characteristic function, the characteristic function of \( Y_{K_m} \) can then be given by

\[ \Phi_{K_m}(i\lambda) = K_m \Lambda(i\lambda) \]

According to Lévy’s theorem, we have

\[
G(y) = \Pr \{ Y_{K_m} > y \} = \frac{1}{2\pi i} \int_{-\infty}^{+\infty} \frac{1}{\zeta} \exp(K_m \Lambda(i\zeta) - i\zeta y) d\zeta
\]

\[
= \frac{1}{2\pi i} \int_{\lambda - i\infty}^{\lambda + i\infty} \frac{1}{z} \exp(K_m \Lambda(z) - zy) dz,
\]

(64)

where \( z = \lambda + i\zeta \), and \( \lambda \) is chosen from the convergence region of this integral. According to the saddle-point approximation method \([21]\), if we let \( z^* \) with \( \Re(z^*) = \lambda^* \) be a solution of the saddle-point equation \( \Lambda'(z) = \frac{y}{K_m} \), then

\[
\frac{1}{2\pi i} \int_{\lambda^* - i\infty}^{\lambda^* + i\infty} \frac{e^{K_m \Lambda(z) - zy}}{z} dz \lesssim \frac{e^{K_m \Lambda(z^*) - z^* y}}{\sqrt{2\pi K_m \Lambda''(\lambda^*)} \lambda^*}.
\]

(65)

Since \( G(y) \) is a real function, we can always choose a real \( z^* \) only if the cumulant-generating function exists. Moreover, \( \lambda^* = \Lambda(\lambda) \) is a conjugate function when \( \lambda \in \mathbb{R} \), then \( \lambda^* \) must be unique. Therefore, the exponentially tight upper bound of \( p_{m,\text{con}}(R_m|D_{K_m,\text{km}}) \) is the tail distribution of \( G(y) \) with \( y > \mathbb{E}\{Y_{K_m}\} \), which is given by

\[
p_{m,\text{con}}(R_m|D_{K_m,\text{km}}) = G(0) \lesssim \frac{1}{\sqrt{2\pi K_m \sigma^2 \lambda^*}} e^{K_m \Lambda(\lambda^*)} = p_{m,\text{upper}}(R_m|D_{K_m,\text{km}}),
\]

(66)

where \( \lambda^* \) is the solution of \( \Lambda'(\lambda) = 0 \), and \( \sigma^2 = \Lambda''(\lambda^*) \). Clearly, we have

\[
\Lambda'(\lambda) = (\alpha_m^* - \ln \gamma) + (1 - \rho_m) \frac{\Gamma'(1 - \lambda, \frac{e^{\alpha_m^*}}{\gamma})}{\Gamma(1 - \lambda, \frac{e^{\alpha_m^*}}{\gamma})} + \rho_m \frac{\Gamma'(1 - \lambda, \frac{1}{\gamma}) - \Gamma'(1 - \lambda, \frac{e^{\alpha_m^*}}{\gamma})}{\Gamma(1 - \lambda, \frac{1}{\gamma}) - \Gamma(1 - \lambda, \frac{e^{\alpha_m^*}}{\gamma})} = 0.
\]

(67)
According to the properties of Gamma function and Meijer’s $G$-function, we have

$$\Gamma'(1 - \lambda, z) = -\Gamma(1 - \lambda, z) \ln z - \frac{1}{G_{2,3}^{3,0}(z)} \begin{pmatrix} 1, & 1, & - \\ 0, & 0, & 1 - \lambda \end{pmatrix}.$$  \hspace{1cm} (68)

Therefore, Eq. (35) can be obtained by plugging Eq. (68) into Eq. (67). For $\sigma^2$, we have

$$\sigma^2 = \Lambda''(\lambda^*) = (1 - \rho_m) \frac{\Gamma''(1 - \lambda^*, \frac{e_{m}^\gamma}{\gamma}) - \left( \frac{\Gamma''(1 - \lambda^*, \frac{e_{m}^\gamma}{\gamma}}{\Gamma(1 - \lambda^*, \frac{e_{m}^\gamma}{\gamma})} \right)^2}{\Gamma(1 - \lambda^*, \frac{1}{\gamma}) - \Gamma(1 - \lambda^*, \frac{e_{m}^\gamma}{\gamma})} - \rho_m \frac{\Gamma''(1 - \lambda^*, \frac{1}{\gamma}) - \Gamma(1 - \lambda^*, \frac{e_{m}^\gamma}{\gamma})}{\Gamma(1 - \lambda^*, \frac{1}{\gamma}) - \Gamma(1 - \lambda^*, \frac{e_{m}^\gamma}{\gamma})}.$$  \hspace{1cm} (69)

According to the properties of Meijer’s $G$-function, we have

$$\frac{\partial}{\partial x} G_{2,3}^{3,0}(z) \begin{pmatrix} 1, & 1, & - \\ 0, & 0, & 1 - \lambda \end{pmatrix} = G_{2,3}^{3,0}(z) \begin{pmatrix} 1, & 1, & - \\ 0, & 0, & 1 - \lambda \end{pmatrix} \ln z + 2G_{3,4}^{4,0}(z) \begin{pmatrix} 1, & 1, & - \\ 0, & 0, & 1 - \lambda \end{pmatrix}.$$  \hspace{1cm} (70)

Eq. (36) can be obtained by plugging Eq. (70) into Eq. (69).

For the tail distribution with $y < \mathbb{E}\{Y_{K_m}\}$, a similar result can be obtained by applying the same process. In this context, Lemma 1 has been established.

**APPENDIX C**

**PROOF OF LEMMA 3**

Let $G(\mathcal{U} \cup \mathcal{A}, \mathcal{E})$ be a bipartite graph with $|\mathcal{U}| = M$ and $|\mathcal{A}| = N$. Clearly, we have $K^{\text{sum}} \leq NL$ and $K_m \leq N$ for $\forall u_m \in \mathcal{U}$. Suppose $\mathcal{E}$ contains at least one maximum $B$-matching such that the only unsaturated vertex is $u_{j,M}$. According to Lemma 2 there must be a subset $\mathcal{X} \subseteq \mathcal{U} \cup \mathcal{A}$ satisfying

$$b(\mathcal{U} \cup \mathcal{A} \setminus \mathcal{X}) + |\mathcal{E}(\mathcal{X})| = \sum_{i=1,i\neq j}^{M} K_{i:M}.$$  \hspace{1cm} (71)

Denote $\mathcal{X}_U = \mathcal{X} \cap \mathcal{U}$ and $\mathcal{X}_A = \mathcal{X} \cap \mathcal{A}$. For any $\mathcal{X}$ which satisfies Eq. (71), there are at most

$$|M - |\mathcal{X}_U||N + |\mathcal{E}(\mathcal{X}_U)||$$  \hspace{1cm} (72)

edges if $\mathcal{X}_U \neq \emptyset$ or

$$|M(N - |\mathcal{X}_A|| + |\mathcal{E}(\mathcal{X}_A)||$$  \hspace{1cm} (73)

edges if $\mathcal{X}_A \neq \emptyset$.
Consider Eq. (72) first. As $|\mathcal{E}(\chi_d)| \leq |\chi_d|N$, the maximum value of Eq. (72) is achieved if and only if $|\chi_d| = 1$. In this case, the only unsaturated vertex $u_{j:M}$ is in $\{u_{i:M}\}_{i=m}^M$. Then, we have $\chi_d = \{u_{m:M}\}$, $\chi_A = \mathcal{A}$, and $|\mathcal{E}(\chi_d)| = K_m - 1$. Therefore, the maximum number of edges in this case is given by

$$\Phi_1 = (M - 1)N + K_m - 1. \quad (74)$$

Consider then Eq. (73). According to Eq. (71), the maximum value of $N - |\chi_A|$ must be $\left\lceil \frac{K_{\text{sum}}}{L} \right\rceil - 1$. Otherwise, there will be no unsaturated vertices in $\mathcal{U}$. In this case, the only unsaturated vertex $u_{j:M}$ is in $\mathcal{U}$. Then, we have $\chi_d = \mathcal{U}$ and $|\mathcal{E}(\chi_A)| = K_{\text{sum}} - L \left( \left\lceil \frac{K_{\text{sum}}}{L} \right\rceil - 1 \right) - 1$. Thus, the maximum number of edges is then given by

$$\Phi_2 = M \left( \left\lceil \frac{K_{\text{sum}}}{L} \right\rceil - 1 \right) + K_{\text{sum}} - L \left( \left\lceil \frac{K_{\text{sum}}}{L} \right\rceil - 1 \right) - 1$$

$$=(M - L) \left( \left\lceil \frac{K_{\text{sum}}}{L} \right\rceil - 1 \right) + K_{\text{sum}} - 1. \quad (75)$$

In this context, Lemma 3 has been established.

**APPENDIX D**

**PROOF OF THEOREM 1**

Consider first the high SNR regime. The first order approximation of $p^\text{out}_{m:M}(R_{m:M})$ is determined by the lowest power term of $p_{m:M}$. According to Lemma 1, $p^\text{con}_{m:M}(R_{m:M}|D_{K_m:M,N})$ has the same decreasing order as $p^{N-K}_{m:M}$ when $\gamma$ tends to infinity. Then, the first part of Lemma 3 indicates that if $\kappa$ Fog-APs are out for $u_{m:M}$ with $\kappa = N - K_m + 1, \ldots, N$, the content outage probability of $u_{m:M}$ for a fixed $\kappa$ is given by

$$\binom{N}{\kappa} p^\text{con}_{m:M}(R_{m:M}|D_{K_m:M,N}) p^\kappa_{m:M} q^{MN-\kappa}_{m:M}$$

$$= \binom{N}{\kappa} p^\text{con}_{m:M}(R_{m:M}|D_{K_m:M,N}) p^\kappa_{m:M} + O(p^N_{m:M}) . \quad (76)$$

Therefore, the first order approximation of $p^\text{out}_{m:M}(R_{m:M})$ can be written by

$$p^\text{out}_{m:M}(R_{m:M}) = \sum_{\kappa=N-K_m+1}^N \binom{N}{\kappa} p^\text{con}_{m:M}(R_{m:M}|D_{K_m:M,N}) p^\kappa_{m:M} + O(p^N_{m:M}) \quad (77)$$

The second part of Lemma 3 indicates that at most $\Phi_2$ edges exist if there is one user in $\mathcal{U}$ is in outage. Recalling Eq. (11), the user $u_m$ will be allocated at least $\eta K_m$ Fog-APs in this case. Therefore, the content outage probability of user $u_m$ is given by

$$p^\text{out}_m(R_m) = \binom{MN}{\Phi_2} p^\text{con}_m(R_m|D_{K_m,\eta K_m}) p^{MN-\Phi_2}_{m} q^{\Phi_2}_{m}$$

$$= \binom{MN}{\Phi_2} p^\text{con}_m(R_m|D_{K_m,\eta K_m}) p^{MN-\Phi_2}_{m} + O(p^{MN-\Phi_2+\eta K_m}) . \quad (78)$$
Therefore, if \( N < MN - \Phi_2 + \eta K_m \), i.e., \( N > \frac{\Phi_2 - \eta K_m}{M-1} \), the first order approximation is given by Eq. (77). On the other hand, if \( N > MN - \Phi_2 + \eta K_m \), i.e., \( N < \frac{\Phi_2 - \eta K_m}{M-1} \), the first order approximation is given by Eq. (78). Finally, if \( N = MN - \Phi_2 + \eta K_m \), i.e., \( N = \frac{\Phi_2 - \eta K_m}{M-1} \), the first order approximation is given by the summation of Eq. (77) and Eq. (78).

Consider then the low SNR regime such that the sample of \( G(K_N;P) \) has a few edges. For a user \( u_m \in \mathcal{U} \), there are two cases that make \( u_m \) unsaturated: 1) There are no \( K_m \) non-outage Fog-APs in \( \mathcal{A} \) for \( u_m \); and 2) There are other users competing for the same Fog-AP with \( u_m \) and it is not saturated by the maximum \( b \)-matching. In the first case, the occurrence probability of this event is given by

\[
\sum_{\kappa=N-K_m+1}^{N} \binom{N}{\kappa} p_m^{\text{con}}(R_m|D_{K_m,N-\kappa}) p_m^{K} q_{m-N-k} = p_m^{N} + N p_m^{\text{con}}(R_m|D_{K_m,1}) p_m^{N-1} q_m + O(q_m),
\]

for \( \gamma \to 0 \). In the second case, there will be at least two edges in the bipartite graph \( G(\mathcal{U} \cup \mathcal{S}, \mathcal{E}) \). One is \( e_{mn} = u_m a_n \), and the other one is \( e_{m'n} u_m a_n \). Assuming that there are only two edges in the sample of this random bipartite graph, i.e., \( \mathcal{E} = \{ e_{mn}, e_{m'n} \} \). In the maximum \( b \)-matching, \( u_m \) or \( u_{m'} \) is chosen with equal probability. The outage probability of \( u_m \) must have the following term

\[
\frac{1}{2} \binom{M}{2} \binom{N}{1} p_m^{\text{con}}(R_m|D_{K_m,1}) p_m^{M-2} q_m^2 = O(q_m),
\]

for \( \gamma \to 0 \). It can be seen that if there are more than two edges in this random bipartite graph, the outage probability of \( u_m \) must have a factor \( q_m^t \) with \( t \geq 3 \). Hence, Eq. (43) holds.

REFERENCES

[1] F. Bonomi, R. Milito, J. Zhu, and S. Addepalli, “Fog computing and its role in the internet of things,” in Proc. ACM MCC ’12. Helsinki, Finland: ACM MCC, Aug. 2012, pp. 13–15.
[2] M. Chiang, “Fog networking: an overview on research opportunities,” arXiv.org, Jan. 2016, arXiv: 1601.00835. [Online]. Available: [http://arxiv.org/abs/1601.00835](http://arxiv.org/abs/1601.00835)
[3] Y. Shi, J. Zhang, K. Letaief, B. Bai, and W. Chen, “Large-scale convex optimization for ultra-dense cloud-RAN,” IEEE Wireless Commun., vol. 22, no. 3, pp. 84–91, Jun. 2015.
[4] M. Maddah-Ali and U. Niesen, “Fundamental limits of caching,” IEEE Trans. Inf. Theory, vol. 60, no. 5, pp. 2856–2867, May 2014.
[5] L. Wang, H. Wu, Z. Han, P. Zhang, and H. V. Poor, “Multi-hop cooperative caching in social IoT using matching theory,” IEEE Trans. Wireless Commun., vol. 17, no. 4, pp. 2127–2145, Apr. 2018. [Online]. Available: [https://ieeexplore.ieee.org/document/8234659/](https://ieeexplore.ieee.org/document/8234659/)
[6] K. Shanmugam, N. Golrezaei, A. G. Dimakis, A. F. Molisch, and G. Caire, “FemtoCaching: wireless content delivery through distributed caching helpers,” IEEE Trans. Inf. Theory, vol. 59, no. 12, pp. 8402–8413, Dec. 2013.
[7] 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.
[8] X. Wang, M. Chen, T. Taleb, A. Ksentini, and V. Leung, “Cache in the air: exploiting content caching and delivery techniques for 5g systems,” IEEE Commun. Mag., vol. 52, no. 2, pp. 131–139, Feb. 2014.

[9] J. Liu, B. Bai, J. Zhang, and K. B. Letaief, “Cache placement in Fog-RANs: from centralized to distributed algorithms,” IEEE Transactions on Wireless Communications, vol. 16, no. 11, pp. 7039–7051, Nov. 2017.

[10] B. Bai, L. Wang, Z. Han, W. Chen, and T. Svensson, “Caching based socially-aware D2d communications in wireless content delivery networks: a hypergraph framework,” IEEE Wireless Communications, vol. 23, no. 4, pp. 74–81, Aug. 2016.

[11] A. G. Dimakis, P. B. Godfrey, Y. Wu, M. J. Wainwright, and K. Ramchandran, “Network coding for distributed storage systems,” IEEE Trans. Inf. Theory, vol. 56, no. 9, pp. 4539–4551, Sep. 2010.

[12] C. Murat and V. T. Paschos, Probabilistic Combinatorial Optimization on Graphs. New York, NY, USA: John Wiley & Sons, 2006.

[13] E. Bjørnson, E. Jorswieck, M. Debbah, and B. Ottersten, “Multiobjective signal processing optimization: the way to balance conflicting metrics in 5g systems,” IEEE Signal Process. Mag., vol. 31, no. 6, pp. 14–23, Nov. 2014.

[14] B. Huang and T. Jebara, “Fast b-matching via sufficient selection belief propagation,” in Proc. Machine Learning Research ’11. Fort Lauderdale, FL, USA: PMLR, Apr. 2011, pp. 361–369.

[15] L. Zheng, “Diversity-multiplexing tradeoff: a comprehensive view of multiple antenna systems,” Ph. D Dissertation, University of California at Berkeley, Berkeley, CA, USA, 2002.

[16] F. Oggier, J.-C. Belfiore, and E. Viterbo, “Cyclic division algebras: a tool for space-time coding,” FnT Communications and Information Theory, vol. 4, no. 1, pp. 1–95, 2007.

[17] D. N. C. Tse and P. Viswanath, Fundamentals of Wireless Communication. New York, NY, USA: Cambridge University Press, 2005.

[18] B. Bai, W. Chen, K. B. Letaief, and Z. Cao, “Outage exponent: a unified performance metric for the parallel fading channel,” IEEE Trans. Inf. Theory, vol. 59, no. 3, pp. 1657–1677, Mar. 2013.

[19] B. Huang and T. Jebara, “Fast b-matching via sufficient selection belief propagation,” in Proc. Machine Learning Research, Fort Lauderdale, FL, USA, 11–13 Apr. 2011, pp. 361–369.

[20] B. Bai, W. Chen, K. B. Letaief, and Z. Cao, “Distributed wrbg matching approach for multiflow two-way d2d networks,” IEEE Transactions on Wireless Communications, vol. 15, no. 4, pp. 2925–2939, April 2016.

[21] R. W. Butler, Saddlepoint Approximations with Applications. New York, NY, USA: Cambridge University Press, 2007.

[22] B. Bai, W. Chen, K. B. Letaief, and Z. Cao, “Conditional outage performance analysis framework for OFDM channels,” in Proc. IEEE ICC ’13. Budapest, Hungary: IEEE ICC, Jun. 2013.

[23] I. S. Gradshteyn and I. M. Ryzhik, Table of Integrals, Series and Products, 7th ed. Burlington, VT, USA: Academic Press, 2007.

[24] A. Schrijver, Combinatorial Optimization: Polyhedra and Efficiency. Berlin, Germany: Springer, 2003.

[25] L. Weng, S. S. Pradhan, and A. Anastasopoulos, “Error exponent regions for Gaussian broadcast and multiple-access channels,” IEEE Trans. Inf. Theory, vol. 54, no. 7, pp. 2919–2942, Jul. 2008.

[26] D. Fudenberg and J. Tirole, Game Theory. Cambridge, MA, USA: The MIT Press, 1991.

[27] A. G. Dimakis, K. Ramchandran, Y. Wu, and C. Suh, “A survey on network codes for distributed storage,” P. IEEE, vol. 99, no. 3, pp. 476–489, Mar. 2011.

[28] B. Bollobás, Random Graphs, 2nd ed. New York, NY, USA: Cambridge University Press, 2001.