Clustered Cell-Free Networking: A Graph Partitioning Approach

Junyuan Wang, *Member, IEEE*, Lin Dai, *Senior Member, IEEE*,

Lu Yang, *Member, IEEE*, and Bo Bai *Senior Member, IEEE*

Abstract

By moving to millimeter wave (mmWave) frequencies, base stations (BSs) will be densely deployed to provide seamless coverage in sixth generation (6G) mobile communication systems, which, unfortunately, leads to severe cell-edge problem. In addition, with massive multiple-input-multiple-output (MIMO) antenna arrays employed at BSs, the beamspace channel is sparse for each user, and thus there is no need to serve all the users in a cell by all the beams therein jointly. Therefore, it is of paramount importance to develop a flexible clustered cell-free networking scheme that can decompose the whole network into a number of weakly interfered small subnetworks operating independently and in parallel. Given a per-user rate constraint for service quality guarantee, this paper aims to maximize the number of decomposed subnetworks so as to reduce the signaling overhead and system complexity as much as possible. By formulating it as a bipartite graph partitioning problem, a rate-constrained network decomposition (RC-NetDecomp) algorithm is proposed, which can smoothly tune the network structure from the current cellular network with simple beam allocation to a fully cooperative network by increasing the required per-user rate. Simulation results demonstrate that the proposed RC-NetDecomp algorithm outperforms existing baselines in terms of average per-user rate, fairness among users and energy efficiency.

Index Terms

Clustered cell-free networking, network decomposition, graph partitioning, spectral clustering, massive multiple-input-multiple-output (MIMO)

I. INTRODUCTION

Mobile communication system has been evolving from its first generation (1G) to the current fifth generation (5G) under the cellular network structure over decades, where each base-station (BS) covers a given area and serves the users within its coverage independently. In order to meet the ever-increasing high data rate requirements, it has been proposed to explore the millimeter wave (mmWave) bands and even the terahertz bands in sixth generation (6G) mobile communication systems [2]. As mmWave/terahertz signals suffer from high path-loss and blockage effect, the coverage of a BS could reduce to a few hundreds or even tens of meters, leading to dense deployment of BSs for the sake of seamless coverage. It could be then expected that a large number of users would be located in the cell-edge areas and thus suffer from severe interference from neighboring BSs under the cellular network structure. Owing to the escalated cell-edge problem and the motivation to serve a user by multiple BSs to avoid signal blockage, we have to rethink whether the current cellular network structure is still suitable for 6G mobile communications.

A. Cell-Free Massive MIMO

An alternative network topology is the large-scale distributed antenna system (DAS) [3]–[8], where a large number of geographically distributed access points (APs)[1] are connected to a central processing unit (CPU) via optical fibers and jointly serve all the users in the system. The DAS has been researched under the cloud radio access network (C-RAN) architecture [9], and was recently redefined as cell-free massive multiple-input-multiple-output (MIMO) in [10] by introducing a simple conjugate beamforming scheme. With conjugate beamforming, each AP estimates the channel state information (CSI) from all the users to it, and then applies conjugate beamforming to serve the users independently. Though without frequent CSI exchange among APs, the number of channels that need to be measured increases linearly with both the number of APs and the number of users, resulting in huge CSI signaling overhead and high signal processing complexity in cell-free massive MIMO systems.

In fact, a user's signal power is mainly contributed by a number of surrounding APs, as the path-loss increases exponentially with the access distance from a user to an AP, indicating that coordinating all the APs to jointly serve all the users is unnecessary. A user-centric *virtual-cell*

---

[1]Note that the AP here can be regarded as a mini BS. In this paper, AP and BS will be used interchangeably.

concept was first proposed in [11], [12] in 2002. Specifically, each user selects a few APs with the highest channel gains to form its own virtual cell and a user is served by the APs within its virtual cell only. Such a user-centric approach is gaining increasing momentum due to the AP densification in 5G systems and beyond [13]–[24]. A user-centric virtual cell is mainly formed in four approaches: 1) each user selects a given number of APs with the best channel qualities or smallest distances [13]–[17]; 2) each AP serves a given number of users with the best channels [18], [19]; 3) each AP serves the users within its coverage of some radius [20]–[22]; and 4) each user selects the APs with the best channels that contribute a given percentage to the overall channel gain [23], [24].

Though simple, the user-centric virtual cells formed by the above methods are usually over-lapped with each other, i.e., an AP could belong to multiple virtual cells, which complicates the downlink transmissions as the virtual cells are coupled with each other by the per-AP power constraint. Moreover, although a user is always served by its surrounding APs in the user-centric approach with cell-edge problem avoided, it still suffers from interference from other virtual cells, which limits the system spectral efficiency. A number of research work was then devoted to the interference management problems in user-centric virtual-cell based networks. For instance, power control algorithms were studied in [15], [19]. A local minimum mean square error (MMSE) receiver was proposed in [16], and a parallel interference cancellation scheme was further designed in [17] to reduce the complexity of uplink signal reception. For downlink transmission, a local partial zero-forcing precoding scheme was proposed in [23]. By introducing orthogonality among virtual cells, [20] and [24] studied the resource block allocation problem and the pilot assignment problem, respectively, based on the graph coloring theory.

B. Clustered Cell-Free Networking

Since the aforementioned approaches optimize the data transmission of virtual cells jointly to manage interference [15]–[17], [19], [23], huge CSI exchange and high computational complexity are inevitable. A more practical network structure to facilitate efficient interference cancellation with tolerable complexity is to decompose the whole network into a number of small subnetworks such that the users strongly interfered with each other are within the same subnetwork and little interference exists among different subnetworks [25]. In this way, efficient interference cancellation techniques can be applied within each subnetwork independently, and the inter-

subnetwork interference could be treated as noise. Such a flexible networking scheme is referred to as *clustered cell-free networking* in this paper.

There are various ways to cluster users and APs into subnetworks. The previously proposed coordinated multi-point (CoMP) transmission [26]–[29] in cellular networks, for instance, can be regarded as an AP-centric clustering scheme, where the APs are first partitioned into multiple clusters and then each user joins the cluster its associated AP belongs to. Various dynamic AP clustering methods were proposed by adopting heuristic methods [30]–[34]. A representative scheme recently proposed in [34] adopted a hierarchical clustering algorithm to group APs based on the minimax linkage criterion [35]. However, with AP-centric clustering algorithms, since the clusters are formed from the AP side only, there could be unlucky users located close to the cluster edge and hence suffer from strong interference from neighboring subnetworks [29].

In order to avoid the cluster-edge problem, user-centric clustering algorithms were proposed [14], [36]–[38]. Specifically, each user forms its own virtual cell by selecting a few nearby APs, and then virtual cells are merged to form subnetworks according to various criteria. For instance, [14] proposed to merge the virtual cells that share at least one AP. As a user is always located at the center of its own virtual cell, there are no cluster-edge users. However, the resulted subnetworks are usually imbalanced with a giant subnetwork as demonstrated in [25], implying that the joint processing complexity and signaling overhead could be still high after clustering.

Note that in the above existing work on clustering users and APs, single-antenna APs are usually assumed. As massive MIMO antenna arrays would be widely deployed in 5G systems and beyond [39], [40], it is of paramount importance to study the clustered cell-free networking problem in a general cell-free wireless network with multiple APs, each equipped with a massive MIMO antenna array. In a massive MIMO system with preformed narrow and directional beams, the beamspace channel from a user to the beams is sparse and a user's beamspace channel is mainly determined by a few beams. Since only users with highly correlated beamspace channels are strongly interfered with each other [41], in order to reduce signaling overhead and system complexity, the beams and users can be grouped into a number of subnetworks with little inter-subnetwork interference. A beam-user clustering algorithm was proposed in [42] in a single-cell setting, which first selects a fixed number of beams for each user with the highest beam gains and then groups strongly interfered users along with their associated beams together for intra-subnetwork interference cancellation. This is essentially a user-centric clustering algorithm similar to that proposed in [14].

In contrast to the aforementioned clustering algorithms which are either AP-centric or user-centric, it was recently proposed in [25] that the network decomposition problem can be formulated as a bipartite graph partitioning problem by modeling the wireless network as a weighted undirected bipartite graph with the edge weight between an AP and a user defined as the corresponding channel gain. The proposed algorithm can significantly reduce inter-subnetwork interference compared to the AP-centric ones, and produce balanced subnetworks in the meantime. There are, nevertheless, a few key issues to be further addressed. First of all, the algorithm proposed in [25] was under the constraint that the sum of the sum-interference-to-sum-signal ratios of all the subnetworks is smaller than some predefined threshold, while no rate performance guarantee of users was provided. In addition, it was required in [25] that each subnetwork contains at least one user and one AP. As a result, all the APs are always involved in data transmission, which may lead to significant waste of energy in future dense mobile communication systems. Intuitively, the APs with no users nearby should be switched into sleep mode for power saving. Furthermore, single-antenna APs were assumed in [25], while massive MIMO antenna arrays would be widely deployed at the APs in 5G/6G mobile communication systems.

C. Our Contributions

In this paper, the clustered cell-free networking problem is studied with new practical considerations including: 1) massive MIMO APs featured with dynamic on-off switch of beams,[2] and 2) a per-user rate constraint imposed for service quality guarantee. By considering a general multi-AP multi-beam system, where L massive MIMO APs are deployed with a large number of preformed fixed beams available at each AP, the cell-free massive MIMO system with single-antenna APs and the single-cell massive MIMO system with fixed beams are included as two special cases.

For clustered cell-free networking, the whole network is decomposed into a number of small subnetworks operating independently and in parallel, with joint processing conducted only inside each subnetwork. Therefore, with more subnetworks, the subnetwork size becomes smaller on average, leading to lower signaling overhead and complexity of joint processing. Yet the intersubnetwork interference may increase, causing rate degradation. It is clear that the number of subnetworks determines a crucial tradeoff between rate performance and complexity/signaling

---

[2]An AP is switched into sleep mode if all the beams belonging to it are turned off.

overhead. In this paper, given a per-user rate constraint for service quality guarantee, we aim to decompose a large-scale network into as many subnetworks as possible, so as to reduce the joint processing complexity and signaling overhead to the most extent.[3]

Specifically, to incorporate massive MIMO APs, a weighted undirected bipartite graph is constructed based on the beams at APs and users. To further enable on-off switch of beams, a novel edge weight model is proposed. Instead of modeling the edge weights between users and beams as the corresponding channel gains as [25] did, in this paper, the edge weights incident to a user are modeled as the corresponding beamspace channel gains normalized by the highest channel gain between this user and all beams. As a result, each user has similar connectivity on the graph and is always incident to an edge with weight 1, indicating that all the users would be treated fairly. By contrast, the edge weights incident to different beams could be of significant difference, as the edge weights of a beam would be very small when it is badly aligned with users or its associated AP is far from users. As a result, isolating the beams with small edge weights, i.e., switching them off, would be prioritized.

Based on the proposed bipartite graph with the novel edge weight model, the clustered cell-free networking problem of maximizing the number of decomposed subnetworks with a per-user rate constraint is transformed to a graph partitioning problem with a graph cut based constraint. A novel rate-constrained network decomposition (RC-NetDecomp) algorithm is then proposed to solve the problem. Simulation results show that the proposed RC-NetDecomp algorithm achieves superior performance over the AP-centric baseline in terms of average per-user rate, user fairness and energy efficiency. Compared to the user-centric baseline, our RC-NetDecomp algorithm enables a flexible fine-grained tuning of the number of decomposed subnetworks and the subnetwork size. Moreover, for a given number of decomposed subnetworks, the proposed RC-NetDecomp algorithm produces more balanced subnetworks with smaller maximum subnetwork size than the user-centric approach, which corroborates its ability to efficiently reduce joint processing complexity and signaling overhead.

---

[3]Note that an alternative problem formulation is to optimize the rate performance for a given constraint of complexity, such as the number of subnetworks or the maximum subnetwork size. Optimizing the rate performance could provide the best services by operating the system at the maximum allowed complexity, while maximizing the number of subnetworks for a given rate constraint could keep the system complexity to a minimum as long as the rate requirement is met. In this paper, we focus on the latter as it aims to minimize the joint processing complexity and signaling overhead as much as possible. It would be of great importance to further study the rate optimization problem for a given constraint of complexity in the future.

The remainder of this paper is organized as follows. Section II introduces the clustered cell-free network model and the rate-constrained network decomposition problem. A graph partitioning based clustered cell-free networking algorithm is proposed in Section III, followed by simulation results in Section IV. Finally, concluding remarks are summarized in Section V.

Throughout this paper, $G = (V, E)$ denotes a graph with vertex set V and edge set E. $\mathbb{E}[\cdot]$ denotes the expectation operator. The superscript T denotes the transpose. $\mathbf{0}_{M \times N}$ denotes an $M \times N$ matrix with all entries zero. $x_{i,j}$ denotes the ith row and jth column entry of matrix $\mathbf{X}$. $|X|$ denotes the cardinality of set X. $X \cap Y$ and $X \cup Y$ denote the intersection and union of set X and set Y, respectively. $X \setminus Y$ denotes the relative complement of set Y in set X. $\emptyset$ denotes the empty set.

II. CLUSTERED CELL-FREE NETWORK MODEL AND PROBLEM FORMULATION

A. Network Model and Bipartite Graph Representation

Consider a general cell-free wireless network with K single-antenna users and L APs. It is assumed that each AP l is equipped with a massive MIMO antenna array with N_l fixed analog beams preformed. The total number of beams N in the network is then given by $N = \sum_{l=1}^{L} N_l$. Let us denote the beam set as $B = \{b_1, b_2, \cdots, b_N\}$ with $|B| = N$, and the user set as $U = \{u_1, u_2, \cdots, u_K\}$ with $|U| = K$. Particularly, with $N_l = 1$, $\forall l = 1, 2, \cdots, L$, there is only one beam available at each AP. In this special case, the term "beam" can be used interchangeably with "AP" and the general cell-free wireless network considered in this paper reduces to a cell-free massive MIMO system that originates from the concepts of DAS and C-RAN. In another special case with $L = 1$, there is only one AP in the network, that is, the network reduces to a single-cell fixed-beam based massive MIMO system.

Since there are two types of nodes in the network, users and beams, and communications only happen between users and beams, let us represent the network by a weighted undirected bipartite graph $G = (V, E)$, where $V = U \cup B$ is the vertex set of users and beams, and $E = \{(u_k, b_n) : \forall u_k \in U, \forall b_n \in B\}$ denotes the set of edges representing the beamspace channels between users and beams. The weight matrix of edges is denoted by $\mathbf{W} \in \mathbb{R}^{K \times N}$, where the kth row and nth column element $w_{k,n}$ is the edge weight between user u_k and beam b_n. Let us rewrite the vertex set as $V = \{v_1, v_2, \cdots, v_{K+N}\}$ with the ith element $v_i = u_i$ if $1 \leq i \leq K$ and
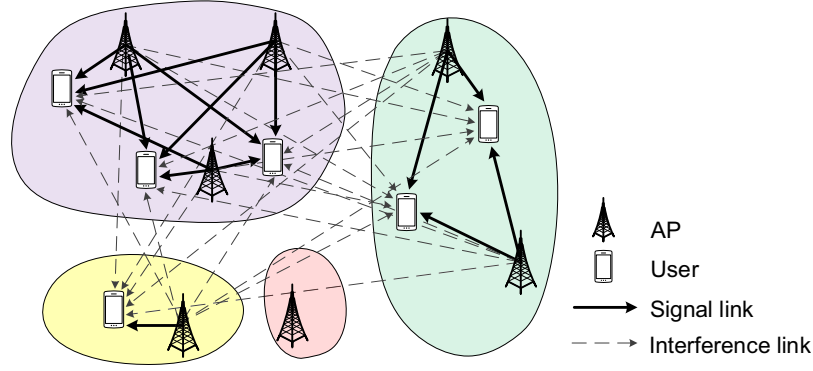
Fig. 1. Illustration of clustered cell-free networking, where only one beam is assumed at each AP for simplicity.

$v_i = b_{i-K}$ if $K + 1 \leq i \leq K + N$. The weighted adjacency matrix $\mathbf{A}$ of graph G is then given by

$$\mathbf{A} = \begin{bmatrix} \mathbf{0}_{K \times K} & \mathbf{W} \\ \mathbf{W}^T & \mathbf{0}_{N \times N} \end{bmatrix}. \tag{1}$$

In a clustered cell-free network, as depicted in Fig. 1, the whole network is decomposed into a number of small subnetworks that operate independently and in parallel, while joint processing can be performed in each subnetwork for intra-subnetwork interference cancellation. It should be noted that the beams available at one AP can be assigned to different subnetworks, which enables flexible cooperation among APs. If a subnetwork contains beams only, these beams will be turned off to save energy. If all the beams belonging to an AP are off, this AP will be switched into sleep mode. Let M denote the number of decomposed subnetworks and C_m denote the set of users and beams in the mth subnetwork. We have

$$\bigcup_{m=1}^{M} C_m = U \cup B = V, \tag{2}$$

and

$$C_m \cap C_{m'} = \emptyset, \ \forall m' \neq m. \tag{3}$$

It can be seen that by modeling the network as a graph, each subnetwork is a subgraph. The clustered cell-free networking problem can be then regarded as a graph partitioning problem. A partition $\mathcal{M}$ of graph G with M subgraphs is denoted as

$$\mathcal{M} = \{C_1, C_2, \cdots, C_M\}. \tag{4}$$

The main notations used in this paper are listed in Table 1 for ease of reading.

TABLE I

MAIN NOTATIONS

K	Number of users
L	Number of APs
N_l	Number of beams at AP l
N	Total number of beams
$B = \{b_1, b_2, \cdots, b_N\}$	Beam set
$U = \{u_1, u_2, \cdots, u_K\}$	User set
$G = (V, E)$	A graph with vertex set V and edge set E
$\mathbf{W}$	Edge weight matrix
$\mathbf{A}$	Weighted adjacency matrix
$\mathbf{D}$	Degree matrix
$\mathbf{L}$	Graph Laplacian matrix
M	Number of decomposed subnetworks
C_m	Set of users and beams in the mth subnetwork
$\mathcal{M} = \{C_1, C_2, \cdots, C_M\}$	A graph partition with M subgraphs
$\text{Cut}(C_m)$	Cut function of subgraph C_m
$R(\mathcal{M})$	Per-user rate achieved under partition $\mathcal{M}$
R_{th}	Minimum per-user rate requirement
μ_k	Received signal-to-interference-plus-noise ratio (SINR) at user u_k
P_t	Transmit power on each beam
σ^2	Variance of additive white Gaussian noise (AWGN)
$g_{k,n}$	Beamspace channel gain coefficient from beam b_n to user u_k
$b_{n_k^*}$	Best beam for user u_k with the highest beamspace channel gain
α	Path-loss exponent

B. Problem Formulation of Rate-Constrained Clustered Cell-Free Networking

In this paper, it is assumed that the intra-subnetwork interference is canceled by some efficient precoding scheme in each subnetwork. To perform such joint processing, the beamspace CSI between beams and users in each subnetwork is required. Therefore, the subnetwork size determines the signaling overhead and complexity of joint processing. With more subnetworks, the joint processing complexity and signaling overhead of each subnetwork is reduced on average. Meanwhile, the rate performance degrades due to the increased inter-subnetwork interference. Under the constraint that the per-user rate must be no smaller than a given threshold R_{th} for service quality guarantee, this paper aims to maximize the number of decomposed subnetworks

M as much as possible, so as to reduce the joint processing complexity and signaling overhead to the most extent. The clustered cell-free networking problem can be formulated as

$$\mathcal{P}1: \max_{\mathcal{M}=\{C_1,C_2,\cdots,C_M\}} M \tag{5}$$

$$\text{s.t.} \quad R(\mathcal{M}) \geq R_{th}, \tag{6}$$

$$\bigcup_{m=1}^{M} C_m = U \cup B, \tag{7}$$

$$C_m \cap C_{m'} = \emptyset, \ \forall m' \neq m, \tag{8}$$

$$C_m \cap B \neq \emptyset, \ \forall m, \tag{9}$$

where (7) and (8) follow the constraints given in (2) and (3). (9) ensures that there is at least one beam in each subnetwork such that all the users in the system can be served. $R(\mathcal{M})$ is the per-user rate achieved under partition $\mathcal{M}$, which is defined as

$$R(\mathcal{M}) \triangleq \frac{1}{K} \sum_{m=1}^{M} \sum_{u_k \in C_m} R_k, \tag{10}$$

where R_k is the achievable rate of user u_k. By normalizing the total system bandwidth into unity and focusing on the spectral efficiency, the achievable rate R_k of user u_k can be expressed as

$$R_k = \log_2 \left(1 + \mu_k\right), \tag{11}$$

where μ_k is the received signal-to-interference-plus-noise ratio (SINR) at user u_k.

Note that the received SINR μ_k is closely dependent on the precoding scheme adopted in each subnetwork. In this paper, instead of focusing on a specific precoding scheme, which may limit the application scenarios of the proposed clustered cell-free networking algorithm, we adopt an approximation of the received SINR μ_k for a general spatial precoding scheme with a large number of beams and/or APs. Specifically, with a large number of geographically distributed APs and/or preformed narrow and directive beams available at each AP, the beamspace channel of a user is highly sparse, i.e., the total channel gain of a user is mainly contributed by its dominant component [8], [45]. As a result, the SINR of reference user $u_k \in C_m$ in the mth subnetwork can be approximated by

$$\mu_k \approx \frac{P_t |g_{k,n_k^*}|^2}{\sum_{b_n \notin C_m, b_n \in B} P_t |g_{k,n}|^2 + \sigma^2}, \tag{12}$$

where $g_{k,n}$ denotes the beamspace channel gain coefficient from beam b_n to user u_k. $b_{n_k^*}$ is the best beam for user u_k with the highest beamspace channel gain, i.e., $n_k^* = \arg\max_{n=1,2,\cdots,N} |g_{k,n}|^2$.

P_t is the transmit power on each beam. σ^2 is the variance of the additive white Gaussian noise (AWGN). Note that the SINR approximation in (12) was shown to be accurate when zero-forcing dirty paper coding (ZF-DPC) is adopted to mitigate the intra-subnetwork interference [43]. In general, (12) serves as a good approximation given that the intra-subnetwork interference is eliminated by some spatial precoding scheme [8], [43], [44].

III. RATE-CONSTRAINED CLUSTERED CELL-FREE NETWORKING

As the formulated optimization problem $\mathcal{P}1$ is difficult to solve due to the mixture of the continuous constraint in (6) and the combinatorial constraints in (7)–(9), in this section, we will first transform problem $\mathcal{P}1$ into a cut-constrained bipartite graph partitioning problem, and then propose a rate-constrained network decomposition algorithm for clustered cell-free networking.

A. Problem Reformulation

Since $f(x) = \log_2(1 + 1/x)$ is a convex function for $x > 0$, a lower-bound of the per-user rate, $R^{lb}(\mathcal{M})$, can be obtained from (10)–(12) as

$$R(\mathcal{M}) \geq R^{lb}(\mathcal{M}) = \log_2\left(1 + \frac{K}{\beta + \sum_{m=1}^{M}\sum_{u_k \in C_m}\sum_{b_n \notin C_m, b_n \in B}\frac{|g_{k,n}|^2}{|g_{k,n_k^*}|^2}}\right), \tag{13}$$

according to Jensen's inequality, where $\beta = \sum_{k=1}^{K}\frac{\sigma^2}{P_t|g_{k,n_k^*}|^2}$.

Note that the cut function $\text{cut}(C_m)$ of a subgraph C_m is defined as the sum weights of the edges between the vertices inside and outside C_m, given by

$$\text{cut}(C_m) = \sum_{v_i \in C_m}\sum_{v_j \in \overline{C}_m} a_{i,j}, \tag{14}$$

where $a_{i,j}$ denotes the ith row and jth column element of adjacency matrix $\mathbf{A}$, and $\overline{C}_m = V \setminus C_m$ is the complement of subset $C_m \subseteq V$. According to (1), $\text{cut}(C_m)$ can be rewritten as

$$\text{cut}(C_m) = \sum_{u_k \in C_m}\sum_{b_n \in \overline{C}_m} w_{k,n} + \sum_{b_n \in C_m}\sum_{u_k \in \overline{C}_m} w_{k,n}. \tag{15}$$

Due to the symmetry of adjacency matrix $\mathbf{A}$, the sumcut function of all the M subgraphs that form a partition of graph G can be obtained as

$$\sum_{m=1}^{M}\text{cut}(C_m) = 2\sum_{m=1}^{M}\sum_{u_k \in C_m}\sum_{b_n \in B, b_n \notin C_m} w_{k,n}. \tag{16}$$

Let us model the edge weight $w_{k,n} \in \mathbf{W}$ between user u_k and beam b_n as

$$w_{k,n} = \frac{|g_{k,n}|^2}{|g_{k,n_k^*}|^2}. \qquad (17)$$

By combining (13) and (16)–(17), the per-user rate lower-bound $R^{lb}(\mathcal{M})$ can be rewritten as

$$R^{lb}(\mathcal{M}) = \log_2 \left(1 + \frac{K}{\beta + \frac{1}{2} \sum_{m=1}^{M} \mathrm{cut}(C_m)} \right). \qquad (18)$$

With $R^{lb}(\mathcal{M}) \geq R_{th}$, the per-user rate constraint in (6) is always satisfied. Therefore, by replacing $R(\mathcal{M})$ with its lower-bound $R^{lb}(\mathcal{M})$, the clustered cell-free networking problem $\mathcal{P}1$ given in (5)–(9) can be transformed to

$$\mathcal{P}2: \max_{\mathcal{M}=\{C_1, C_2, \cdots, C_M\}} M \qquad (19)$$

$$\text{s.t.} \quad \sum_{m=1}^{M} \mathrm{cut}(C_m) \leq \frac{2K}{2^{R_{th}} - 1} - 2\beta, \qquad (20)$$

$$\bigcup_{m=1}^{M} C_m = U \cup B, \qquad (21)$$

$$C_m \cap C_{m'} = \emptyset, \ \forall m' \neq m, \qquad (22)$$

$$C_m \cap B \neq \emptyset, \ \forall m. \qquad (23)$$

Remark: Different from [25] where APs are always activated, the clustered cell-free networking problem formulated in this paper could not only produce balanced subnetworks to reduce system complexity and signaling overhead efficiently, but also switch off beams that are badly aligned with users or associated with APs far from users to save energy. Specifically, by defining the edge weight between a user and a beam as the normalized beamspace channel gain, a beam with no users close to its main direction or its associated AP would have small edge weights, and thus lead to small sumcut when forming a subnetwork alone. With the aim of maximizing the total number of subnetworks, isolating such beams would be prioritized, which facilitates the on-off switch of beams for energy saving. From the perspective of users, the maximum edge weight incident to each user is always 1, indicating balanced connectivity of users on the constructed graph G, and thus balanced sizes of the subnetworks containing both users and beams.

B. Rate-Constrained Network Decomposition Algorithm

To solve the clustered cell-free networking problem $\mathcal{P}2$ given in (19)–(23), let us take the double-loop approach in [25]. Specifically, in the inner loop, an M-way graph partitioning

problem is solved to find the optimal partition $\mathcal{M}^*_{|M}$ that minimizes the sumcut $\sum_{m=1}^{M} \mathrm{cut}(C_{m|M})$ given the number of subgraphs M. In the outer loop, the maximum number of subgraphs M^* is searched as long as its corresponding M-way partition $\mathcal{M}^*_{|M^*}$ satisfies the sumcut constraint in (20).

1) M-Way Partitioning: In the inner loop, we aim at finding the optimal partition $\mathcal{M}^*_{|M} = \left\{ C^*_{1|M}, C^*_{2|M}, \cdots, C^*_{M|M} \right\}$ to minimize the sumcut, which can be presented as

$$\mathcal{P}3: \min_{\mathcal{M}_{|M} = \left\{ C_{1|M}, C_{2|M}, \cdots, C_{M|M} \right\}} \sum_{m=1}^{M} \mathrm{cut}(C_{m|M}) \tag{24}$$

$$\text{s.t.} \quad \bigcup_{m=1}^{M} C_{m|M} = U \cup B, \tag{25}$$

$$C_{m|M} \cap C_{m'|M} = \emptyset, \ \forall m' \neq m, \tag{26}$$

$$C_{m|M} \cap B \neq \emptyset, \ \forall m. \tag{27}$$

In order to solve problem $\mathcal{P}3$, let us first convert the bipartite graph partitioning problem into a graph partitioning problem.

By merging each beam vertex b_i with the user vertices with edge weight of 1 between them to form a new meganode $\tilde{v}_i$ as

$$\tilde{v}_i = \left\{ b_i \cup \{ u_k : n^*_k = b_i, \forall u_k \in U \} \right\}, \tag{28}$$

a new graph $\tilde{G} = (\tilde{\mathcal{V}}, \tilde{E})$ of N meganodes can be constructed, where $\tilde{\mathcal{V}} = \{ \tilde{v}_1, \tilde{v}_2, \cdots, \tilde{v}_N \}$ is the meganode set. $\tilde{E} = \left\{ (\tilde{v}_i, \tilde{v}_j) : \forall \tilde{v}_i \in \tilde{\mathcal{V}}, \forall \tilde{v}_j \in \tilde{\mathcal{V}}, i \neq j \right\}$ is the set of edges, where the edge weight $\tilde{w}_{i,j}$ between meganode $\tilde{v}_i$ and meganode $\tilde{v}_j$ is defined as the sum weight of the edges in the original bipartite graph G connecting the vertices inside meganode $\tilde{v}_i$ and the vertices inside meganode $\tilde{v}_j$, that is,

$$\tilde{w}_{i,j} = \begin{cases} \sum_{u_k \in \tilde{v}_i} w_{k,j} + \sum_{u_k \in \tilde{v}_j} w_{k,i}, & \text{if } i \neq j; \\ 0, & \text{otherwise.} \end{cases} \tag{29}$$

A toy example is presented in Fig. 2 to illustrate the construction of the new graph $\tilde{G} = (\tilde{\mathcal{V}}, \tilde{E})$, where beam 1 is merged with users 1 and 2 to form meganode 1, and beam 3 and user 3 are merged as meganode 3. Appendix A shows that the optimization problem $\mathcal{P}3$ given in (24)–(27)
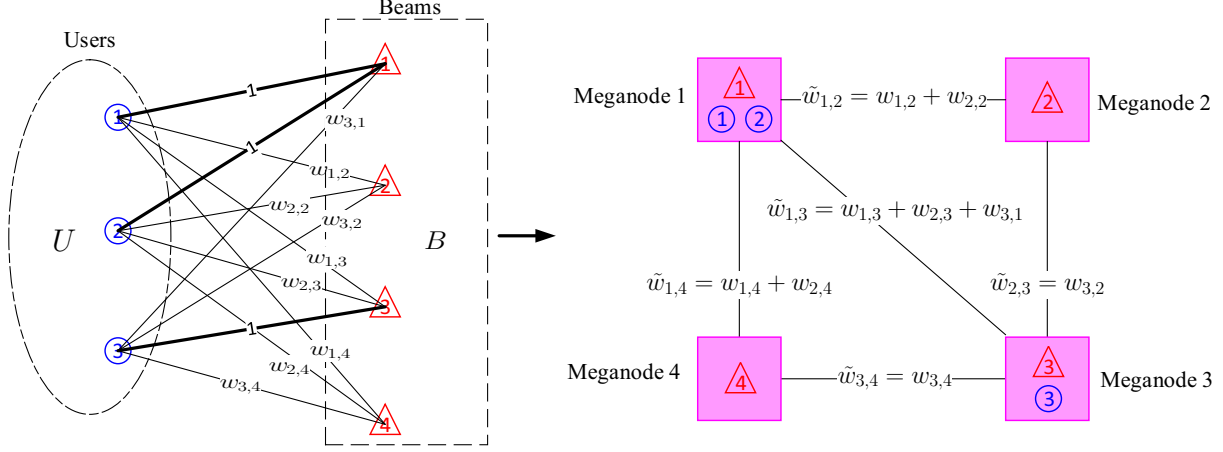
Fig. 2. A toy example of 3 users and 4 beams to illustrate the construction of graph $\tilde{G} = (\tilde{V}, \tilde{E})$. In the bipartite graph on the left-hand side, the edges with weight 1 are plotted in thick lines.

is equivalent to the following graph partitioning problem on graph $\tilde{G} = (\tilde{\mathcal{V}}, \tilde{E})$:

$$\mathcal{P}4 : \quad \min_{\tilde{\mathcal{M}}_{|M} = \left\{ \tilde{\mathcal{C}}_{1|M}, \tilde{\mathcal{C}}_{2|M}, \cdots, \tilde{\mathcal{C}}_{M|M} \right\}} \quad \sum_{m=1}^{M} \mathrm{cut}(\tilde{\mathcal{C}}_{m|M}) \tag{30}$$

$$\text{s.t.} \quad \bigcup_{m=1}^{M} \tilde{\mathcal{C}}_{m|M} = \tilde{\mathcal{V}}, \tag{31}$$

$$\tilde{\mathcal{C}}_{m|M} \cap \tilde{\mathcal{C}}_{m'|M} = \emptyset, \ \forall m' \neq m, \tag{32}$$

where $\tilde{\mathcal{M}}_{|M} = \left\{ \tilde{\mathcal{C}}_{1|M}, \tilde{\mathcal{C}}_{2|M}, \cdots, \tilde{\mathcal{C}}_{M|M} \right\}$ denotes a partition of graph $\tilde{G}$ with M subgraphs.

It can be seen that the optimization problem $\mathcal{P}4$ is a canonical mincut problem, which can be efficiently solved by a spectral clustering method [46]. Specifically, let

$$\mathbf{L} = \mathbf{D} - \tilde{\mathbf{W}} \tag{33}$$

denote the graph Laplacian matrix of graph $\tilde{G}$, where $\tilde{\mathbf{W}}$ is the weighted adjacency matrix with each element $\tilde{w}_{i,j}$ defined in (29). $\mathbf{D}$ is the degree matrix given by

$$\mathbf{D} = diag(d_1, d_2, \cdots, d_N), \tag{34}$$

where d_i is the degree of meganode $\tilde{v}_i$ with

$$d_i = \sum_{j=1}^{N} \tilde{w}_{i,j}. \tag{35}$$

By computing the eigenvectors $\mathbf{p}_1, \mathbf{p}_2, \cdots, \mathbf{p}_M$ of graph Laplacian matrix $\mathbf{L}$ corresponding to the M smallest eigenvalues with $\mathbf{p}_m \in \mathbb{R}^{N \times 1}$, $m = 1, 2, \cdots, M$, a new matrix

$$\mathbf{Y} = [\mathbf{p}_1, \mathbf{p}_2, \cdots, \mathbf{p}_M] \tag{36}$$

can be formed. Let $\mathbf{y}_i^r \in \mathbb{R}^{1 \times M}$ be the ith row vector of $\mathbf{Y}$. By clustering the N vectors $\mathbf{y}_1^r, \mathbf{y}_2^r, \cdots, \mathbf{y}_N^r$ into M clusters $\hat{C}_1, \hat{C}_2, \cdots, \hat{C}_M$ with the k-means algorithm, the optimal $\tilde{\mathcal{C}}_{m|M}^*$ for any $m = 1, 2, \cdots, M$ can be obtained as

$$\tilde{\mathcal{C}}_{m|M}^* = \left\{ \tilde{v}_i : \mathbf{y}_i^r \in \hat{C}_m, i = 1, 2, \cdots, N \right\}. \tag{37}$$

By combining (28) and (37), the optimal partition $\mathcal{M}_{|M}^* = \left\{ C_{1|M}^*, C_{2|M}^*, \cdots, C_{M|M}^* \right\}$ can be obtained with

$$C_{m|M}^* = \left\{ b_i : \tilde{v}_i \in \tilde{\mathcal{C}}_{m|M}^* \right\} \cup \left\{ u_k : u_k \in \tilde{v}_i, \tilde{v}_i \in \tilde{\mathcal{C}}_{m|M}^* \right\}, m = 1, 2, \cdots, M. \tag{38}$$

2) Maximization of M: Based on the solution of the inner-loop M-way partitioning problem, $\mathcal{M}_{|M}^* = \left\{ C_{1|M}^*, C_{2|M}^*, \cdots, C_{M|M}^* \right\}$, the outer-loop problem can be written as

$$\mathcal{P}5: \quad \max \quad M \tag{39}$$

$$\text{s.t.} \quad \sum_{m=1}^{M} \text{cut}(C_{m|M}^*) \leq \frac{2K}{2^{R_{th}} - 1} - 2\beta. \tag{40}$$

As the exhaustive search of the optimal number of decomposed subnetworks M^* would lead to high computational complexity with a massive number of beams N, based on the monotonic property of the mincut function $\sum_{m=1}^{M} \text{cut}(C_{m|M}^*)$ shown in Theorem 1, M^* can be obtained by adopting a binary search approach.

Theorem 1. *The mincut function $\sum_{m=1}^{M} cut(C_{m|M}^*)$ satisfies*

$$\sum_{m=1}^{M} cut(C_{m|M}^*) \leq \sum_{m=1}^{M+1} cut(C_{m|M+1}^*), \tag{41}$$

for any $M = 1, 2, \cdots, N - 1$.

Proof: See Appendix A. ∎

By combining the spectral clustering algorithm solving the M-way partitioning subproblem in the inner loop and the binary search algorithm solving the maximization of M subproblem in the outer loop, a rate-constrained network decomposition (RC-NetDecomp) algorithm is proposed for clustered cell-free networking with detailed steps summarized in Algorithm 1.

Algorithm 1 RC-NetDecomp

1: **Input:** Minimum per-user rate requirement R_{th}, weight matrix $\mathbf{W}$.

2: **Initialization:** $min = 1$, $max = N + 1$, $\mathcal{M}^* = \{V\}$.

3: **while** $min < max - 1$ **do**

4:     $M = \lfloor \frac{min+max}{2} \rfloor$;

5:     Compute $\tilde{\mathbf{W}}$, $\mathbf{L}$ and $\mathbf{Y}$ based on (29) and (33)–(36);

6:     Run the k-means algorithm to cluster row vectors of $\mathbf{Y}$, $\mathbf{y}_1^r, \mathbf{y}_2^r, \cdots, \mathbf{y}_N^c$, into M clusters $\hat{C}_1, \hat{C}_2, \cdots, \hat{C}_M$;

7:     Compute $\mathcal{M}_{|M}^* = \left\{ C_{1|M}^*, C_{2|M}^*, \cdots, C_{M|M}^* \right\}$ according to (37) and (38);

8:     **if** $\sum_{m=1}^M \mathrm{cut}(C_{m|M}^*) \leq \frac{2K}{2^{R_{th}}-1} - 2\beta$ **then**

9:         $min = M$, $\mathcal{M}^* = \mathcal{M}_{|M}^*$;

10:    **else**

11:        $max = M$;

12:    **end if**

13: **end while**

14: **Output:** $\mathcal{M}^*$.

C. Complexity Analysis

Since a binary search method is adopted in the outer loop to find the maximum number of subnetworks M^* on graph $\tilde{G}$, where the number of vertices is N, the computational complexity of the outer-loop binary search is $O(\log N)$. For the inner-loop that partitions graph $\tilde{G}$ into M subgraphs by adopting a spectral clustering method, the complexity is mainly determined by the eigenvalue decomposition of the Laplacian matrix $\mathbf{L}$ of graph $\tilde{G}$, which has the computational complexity of $O(N^3)$. The overall computational complexity of the proposed RC-NetDecomp algorithm is then $O(N^3 \log N)$. Note that by constructing the new graph $\tilde{G}$, the number of vertices in the graph partitioning problem reduces from $K + N$ to N, leading to lower computational complexity than partitioning the original bipartite graph G directly as a byproduct.

IV. SIMULATION RESULTS

In this section, simulation results are presented to demonstrate the performance of the proposed RC-NetDecomp algorithm for clustered cell-free networking. A cell-free wireless network is considered, where users and APs are assumed to be randomly distributed by following an

independent and identical uniform distribution. Note that decomposing a network according to instantaneous channel gains leads to rapidly varying network structure, which is difficult to be applied in practical systems. To this end, let us only consider the path-loss and beamforming gain when performing the RC-NetDecomp algorithm. The beamspace channel gain from beam b_n to user u_k is then set as $|g_{k,n}|^2 = d_{k,n}^{-\alpha} \cdot D_{k,n}$, where $d_{k,n}$ is the distance between user u_k and the AP to which beam b_n belongs, and α is the path-loss exponent. $D_{k,n}$ denotes the power gain of beam b_n at user u_k. The corresponding edge weight $w_{k,n}$ given in (17) can be then written as

$$w_{k,n} = \frac{d_{k,n}^{-\alpha} \cdot D_{k,n}}{d_{k,n_k^*}^{-\alpha} \cdot D_{k,n_k^*}}, \tag{42}$$

where $n_k^* = \arg\max_{n=1,2,\cdots,N} d_{k,n}^{-\alpha} D_{k,n}$. After performing the RC-NetDecomp algorithm, zero-forcing beamforming (ZFBF) is applied in each subnetwork. Since ZFBF requires that the number of beams is no less than the number of users, the rate of each user is zero if there are more users than beams in a subnetwork.

As discussed in Section I, there are two kinds of algorithms in the literature that could be deployed for clustered cell-free networking.

- User-centric clustering algorithms: Subnetworks are formed by merging user-centric virtual cells. A representative user-centric clustering algorithm was proposed in [14], where each user first chooses S APs with the highest channel gains, and then the overlapped virtual cells are merged.

- AP-centric clustering algorithms: APs are first clustered and then each user joins the subnetwork containing its associated AP. A representative AP-centric clustering algorithm was proposed in [34], where a hierarchical clustering algorithm with minimax linkage criterion was adopted to cluster APs.

To demonstrate the performance gains of the proposed RC-NetDecomp algorithm, the AP-centric algorithm in [34] and the user-centric algorithm in [14] are adopted as benchmarks. Note that the RC-NetDecomp algorithm is proposed for a general cell-free wireless network with L APs and N_l beams available at AP l, which reduces to the special cases of cell-free massive MIMO with $N_l = 1$, $\forall l = 1, 2, \cdots, L$, and single-cell massive MIMO with $L = 1$. In the following, we will compare the proposed RC-NetDecomp algorithm against the aforementioned two baselines in these two special cases. Performance metrics of interest include the average per-user rate $\bar{R} \triangleq \mathbb{E}_{\{|g_{k,n}|^2\}_{u_k \in U, b_n \in B}}[R]$, the average minimum rate of users $\bar{R}_{min} \triangleq \mathbb{E}_{\{|g_{k,n}|^2\}_{u_k \in U, b_n \in B}}[\min_{u_k \in U} R_k]$, the average variance of users' rates $\bar{R}_{var} \triangleq$

(a) $R_{th} = 3.5$ bit/s/Hz   (b) $R_{th} = 2$ bit/s/Hz   (c) $R_{th} = 0$
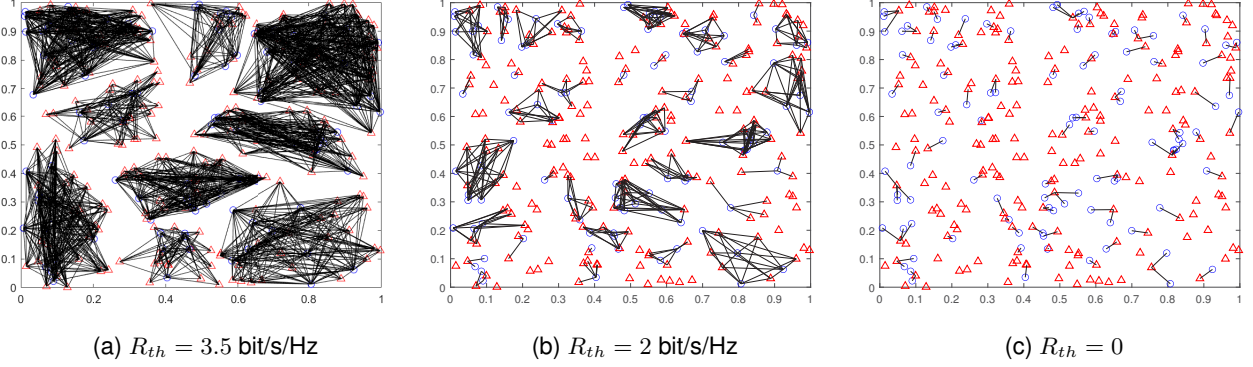
Fig. 3. Decomposed subnetworks of a randomly generated cell-free massive MIMO system by adopting the proposed RC-NetDecomp algorithm under various required per-user rate R_{th}. "$\circ$" represents a user and "$\triangle$" represents an AP. $L = 200$, $K = 100$, $\alpha = 4$.

$\mathbb{E}_{\{|g_{k,n}|^2\}_{u_k \in U, b_n \in B}} \left[\frac{1}{K} \sum_{u_k \in U} (R_k - R)^2 \right]$, the joint processing complexity indicated by the average number of decomposed subnetworks $\bar{M}^* \triangleq \mathbb{E}_{\{|g_{k,n}|^2\}_{u_k \in U, b_n \in B}}[M^*]$ along with the average maximum subnetwork size $\overline{|C|}_{max} \triangleq \mathbb{E}_{\{|g_{k,n}|^2\}_{u_k \in U, b_n \in B}} \left[\max_{m=1,2,\cdots,M^*} |C^*_{m|M^*}| \right]$, and the system energy efficiency indicated by the average percentage of beams switched off $\bar{P}_{off} \triangleq \mathbb{E}_{\{|g_{k,n}|^2\}_{u_k \in U, b_n \in B}} \left[\frac{1}{N} \sum_{m=1, C^*_{m|M^*} \cap U = \emptyset}^{M^*} |C^*_{m|M^*}| \right]$, all of which are obtained by averaging over 500 random realizations of users' and APs' locations.

A. Special Case of Cell-Free Massive MIMO with $N_l = 1$

With the number of beams $N_l = 1$ for any $l = 1, 2, \cdots, L$, each AP is equipped with a single omnidirectional antenna and the general cell-free wireless network considered in this paper reduces to a cell-free massive MIMO system. In such a special case, a beam is equivalent to an AP, and there is no beamforming gain, i.e., $D_{k,n} = 1$, $\forall b_n \in B, \forall u_k \in U$. As a result, the edge weight in (42) reduces to $w_{k,n} = d_{k,n}^{-\alpha}/d_{k,n_k^*}^{-\alpha}$, where d_{k,n_k^*} equals the smallest distance from user u_k to the L APs, i.e., $d_{k,n_k^*} = \min_{n=1,2,\cdots,L} d_{k,n}$.

1) Decomposition Results: Fig. 3 illustrates the decomposition results of a cell-free massive MIMO system with a randomly generated topology by adopting the proposed RC-NetDecomp algorithm under various values of the per-user rate constraint R_{th}. It can be seen in Fig. 3(a) that with a relatively high $R_{th} = 3.5$ bit/s/Hz, the whole network is decomposed into 9 subnetworks, and all the APs are involved in joint transmission. Intuitively, when a sufficiently high data rate is required, all the APs should be activated to contribute to data transmission. By comparing Figs. 3(a)–3(c), it can be found that as the per-user rate constraint R_{th} decreases, more and
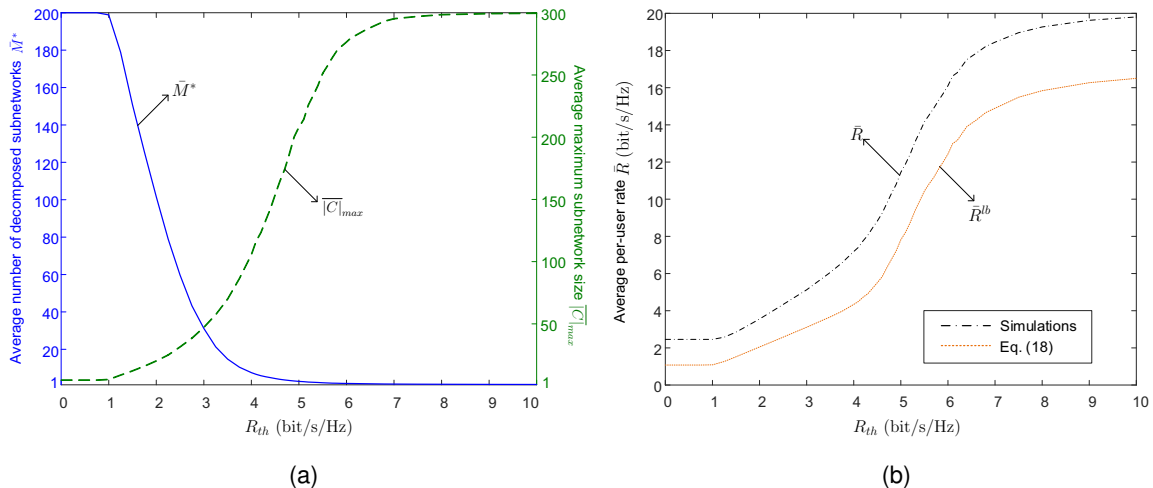
Fig. 4. (a) Average number of decomposed subnetworks $\bar{M}^*$ and average maximum subnetwork size $\overline{|C|}_{max}$, and (b) average per-user rate $\bar{R}$ versus the per-user rate constraint R_{th} with the proposed RC-NetDecomp algorithm. $L = 200$, $K = 100$, $P_t/\sigma^2 = 0$dB, $\alpha = 4$.

more APs that are relatively far away from users become isolated, which could be switched into sleep mode to save energy. This corroborates that our clustered cell-free networking problem has successfully incorporated the sleep mode operation of APs, which is especially important for future dense wireless networks with a larger number of APs. In the case without rate performance guarantee, i.e., $R_{th} = 0$, each user is simply associated to its closest AP only, as shown in Fig. 3(c), i.e., the current cellular network structure.

To have a closer look at the transition of the network structure by increasing the per-user rate constraint R_{th}, Fig. 4(a) presents the average number of decomposed subnetworks $\bar{M}^*$ and the average maximum subnetwork size $\overline{|C|}_{max}$ with the RC-NetDecomp algorithm. It is shown in Fig. 4(a) that as the per-user rate requirement R_{th} increases from 0 to 10 bit/s/Hz, the average number of decomposed subnetworks $\bar{M}^*$ gradually decreases from 200 to 1. Specifically, when R_{th} is smaller than the lower-bound of the per-user rate achieved under the cellular structure, M^* remains at the number of APs $L = 200$. As R_{th} increases beyond that, joint processing of APs would be required for interference cancellation to improve the per-user rate performance. As a result, the number of decomposed subnetworks M^* decreases with R_{th} until $M^* = 1$, i.e., all the APs jointly serve all the users in the network. For the same reason, the average maximum subnetwork size $\overline{|C|}_{max}$ increases with the threshold R_{th}, indicating increased joint processing complexity and signaling overhead. It can be clearly seen that with the proposed RC-NetDecomp
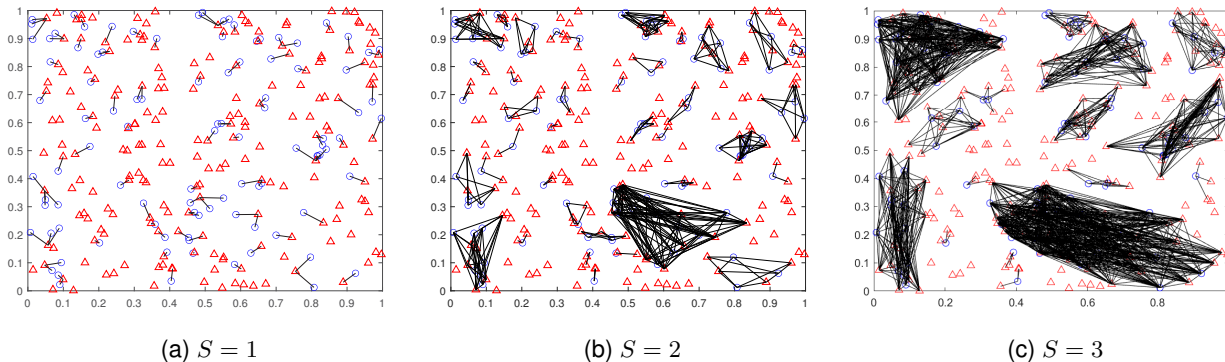
Fig. 5. Decomposed subnetworks of the randomly generated cell-free massive MIMO system shown in Fig. 3 by adopting the user-centric clustering algorithm in [14]. "∘" represents a user and "△" represents an AP. $L = 200$, $K = 100$, $\alpha = 4$.

algorithm, the network structure varies smoothly from the current cellular network to the fully cooperative network by increasing the per-user rate constraint R_{th}.

Fig. 4(b) further shows the corresponding average per-user rate $\bar{R}$ achieved and the average of the per-user rate lower-bound given in (13), $\bar{R}^{lb}$. Both of them are well above the per-user rate constraint R_{th}, indicating that the proposed RC-NetDecomp algorithm can always meet the per-user rate requirement. Since the signaling overhead and joint processing complexity inside each subnetwork increases with the subnetwork size, it can be concluded from Fig. 4 that the required per-user rate R_{th} determines a crucial rate-complexity/signaling overhead tradeoff.

2) Comparison with User-Centric Clustering: With the user-centric clustering algorithm in [14], Fig. 5 presents the network decomposition results under the same topology presented in Fig. 3 by varying the number of APs chosen by each user, S. It can be clearly seen that with $S = 1$, the network structure is the same as that in Fig. 3(c), i.e., the cellular network. As the number of APs chosen by each user, S, increases, more and more users are grouped together for joint processing. However, a giant subnetwork appears when $S = 3$, implying that the system complexity and signaling overhead could be still high after network decomposition.

Fig. 6 further presents the average number of decomposed subnetworks[4] $\bar{M}^*$ and the average maximum subnetwork size $\overline{|C|}_{max}$ by increasing the number of APs chosen by each user, S, from 1 to 200. A similar network structure transition can be observed by tuning S. However, only a small number of options are available as marked in Fig. 6, since S needs to be an integer. Moreover, the average number of decomposed subnetworks $\bar{M}^*$ drops quickly as S

[4]To align with the assumptions in this paper, an isolated AP is counted as a single subnetwork.
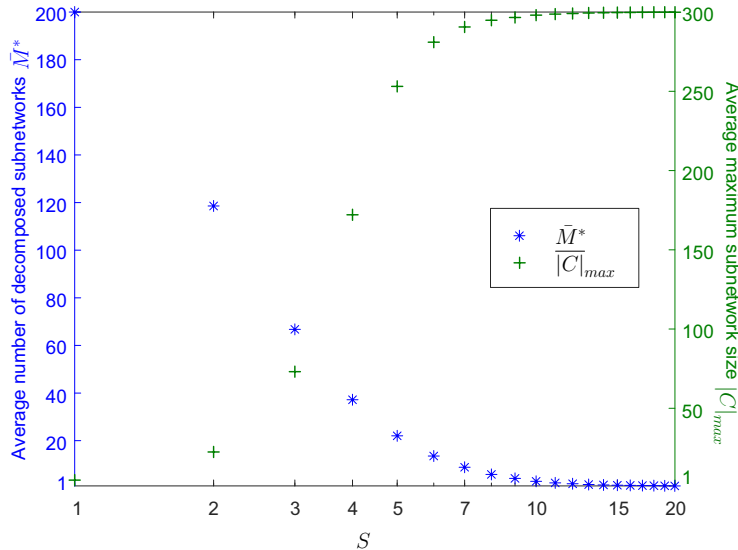
Fig. 6. Average number of decomposed subnetworks $\bar{M}^*$ and average maximum subnetwork size $\overline{|C|}_{max}$ versus the number of APs chosen by each user, S, with the user-centric clustering algorithm in [14]. $L = 200$, $K = 100$, $\alpha = 4$.

increases. By comparing Fig. 6 with Fig. 4(a), we can see that with the same $\bar{M}^*$, the average maximum subnetwork size $\overline{|C|}_{max}$ with the user-centric clustering algorithm in [14] is much larger than that with the RC-NetDecomp algorithm. For instance, with $\bar{M}^* = 67$ (i.e., $S = 3$ in Fig. 6), $\overline{|C|}_{max} = 73$ with the user-centric clustering algorithm. By contrast, with the RC-NetDecomp algorithm, the average maximum subnetwork size $\overline{|C|}_{max}$ is significantly reduced to 30, indicating much lower complexity and signaling overhead for joint processing. It can be concluded that compared to the user-centric clustering benchmark, the proposed RC-NetDecomp algorithm enables a much finer-grained tuning of subnetwork size and more efficient reduction of system complexity and signaling overhead.

3) Comparison with AP-Centric Clustering: With the AP-centric clustering algorithm proposed in [34], Fig. 7 presents the snapshots of the decomposed subnetworks under the same topology shown in Fig. 3 with various numbers of decomposed subnetworks M. It can be clearly seen from Fig. 7 that the clustered cell-free network structures with the AP-centric clustering algorithm in [34] are similar to those with the proposed RC-NetDecomp algorithm, indicating that the AP-centric approach can also enable flexible network decomposition. However, there could be a large number of users located at the subnetwork edge, who would suffer from strong inter-subnetwork interference. By contrast, with the proposed RC-NetDecomp algorithm, it is

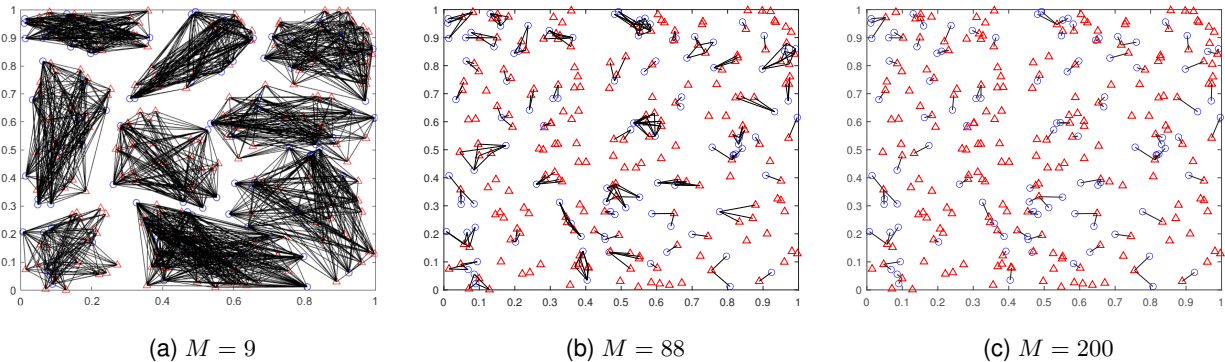(a) $M = 9$ (b) $M = 88$ (c) $M = 200$

Fig. 7. Decomposed subnetworks of the randomly generated cell-free massive MIMO system shown in Fig. 3 by adopting the AP-centric clustering algorithm in [34]. "○" represents a user and "△" represents an AP. $L = 200$, $K = 100$, $\alpha = 4$.

the APs that are usually located at the subnetwork edge as the edge weights on graph G are modeled in a user-centric way. As a result, higher per-user rate could be expected with the proposed RC-NetDecomp algorithm.

Figs. 8(a)–8(c) present the average per-user rate $\bar{R}$, the average minimum rate of users $\bar{R}_{min}$ and the average variance of users' rates $\bar{R}_{var}$ with the proposed RC-NetDecomp algorithm and the AP-centric clustering algorithm in [34]. For the proposed RC-NetDecomp algorithm, it can be observed that both the average per-user rate $\bar{R}$ and the average minimum rate $\bar{R}_{min}$ decrease as the number of subnetworks M increases, yet the average variance of users' rates $\bar{R}_{min}$ first decreases and then increases when M exceeds 80. To understand why the average rate variance increases with M when M is large, note from Fig. 8(b) that the average minimum rate $\bar{R}_{min}$ drops to zero at $M = 80$. With M further increasing, more and more users are in outage, thus leading to higher average variance of users' rates, as observed in Fig. 8(c).

For the comparision with the AP-centric algorithm in [34], it can be seen from Figs. 8(a)–8(c) that the proposed RC-NetDecomp algorithm achieves higher average per-user rate $\bar{R}$, higher average minimum rate of users $\bar{R}_{min}$ and smaller average variance of users' rates $\bar{R}_{var}$, where the latter two indicate that more uniform rate performance could be expected by adopting the proposed RC-NetDecomp algorithm. It can be concluded that compared to the AP-centric clustering benchmark, our proposed RC-NetDecomp algorithm improves not only per-user rate performance but also user fairness. The comparison highlights the importance of clustering based on the locations of both APs and users.

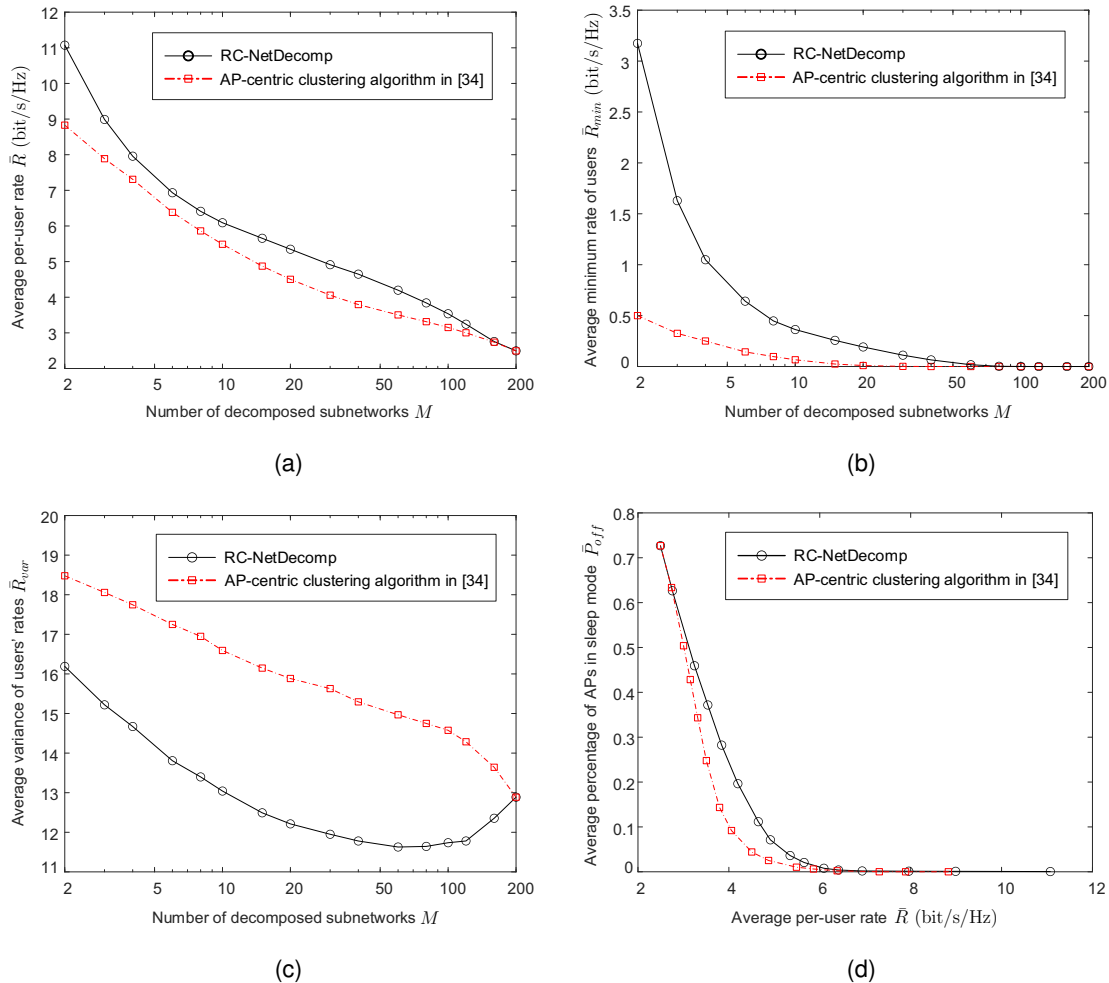In addition, Fig. 8(d) presents the average percentage of APs in sleep mode $\bar{P}_{off}$ with both

Fig. 8. (a) Average per-user rate $\bar{R}$, (b) average minimum rate of users $\bar{R}_{min}$, (c) average variance of users' rates $\bar{R}_{var}$ and (d) average percentage of APs in sleep mode $\bar{P}_{off}$ with the proposed RC-NetDecomp algorithm and the AP-centric clustering algorithm in [34]. $L = 200$, $K = 100$, $P_t/\sigma^2 = 0$dB, $\alpha = 4$.

RC-NetDecomp and the AP-centric clustering algorithm in [34]. We can see that given the same average per-user rate $\bar{R}$, the average percentage of APs in sleep mode $\bar{P}_{off}$ with the proposed RC-NetDecomp algorithm is much higher than that with the AP-centric clustering algorithm in [34] when the average per-user rate $\bar{R}$ is between 3 bit/s/Hz and 6 bit/s/Hz, which indicates significant improvement in energy efficiency by adopting our RC-NetDecomp algorithm.

B. Special Case of Single-Cell Massive MIMO with $L = 1$

With the number of APs $L = 1$, all the N beams are available at the same AP and the general cell-free wireless network considered in this paper reduces to a fixed-beam based single-
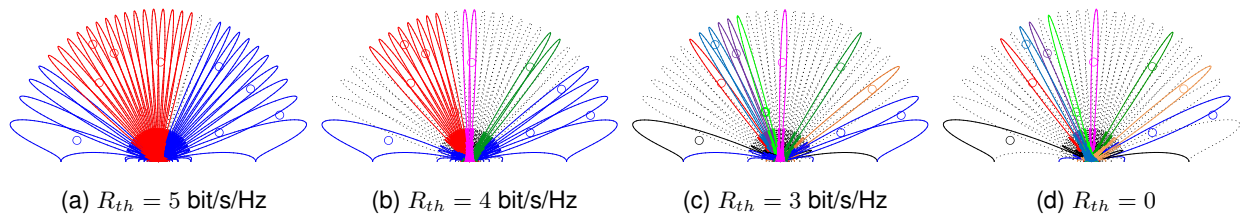
(a) $R_{th} = 5$ bit/s/Hz   (b) $R_{th} = 4$ bit/s/Hz   (c) $R_{th} = 3$ bit/s/Hz   (d) $R_{th} = 0$

Fig. 9. Decomposition results of a fixed-beam based single-cell massive MIMO system by adopting the proposed RC-NetDecomp algorithm. "∘" represents a user. Users and beams in the same color are in the same subnetwork. Inactive beams are drawn in dashed lines. $N = 32$, $K = 10$.

cell massive MIMO system. In such a special case, the distance $d_{k,n}$ is independent of the beam index b_n. As a result, the edge weight given in (42) reduces to $w_{k,n} = D_{k,n}/D_{k,n_k^*}$, where D_{k,n_k^*} is equal to the largest beam gain of user u_k, i.e., $D_{k,n_k^*} = \max_{n=1,2,\cdots N} D_{k,n}$. For simplicity, let us assume that the AP is located at the center of a circular cell and a linear antenna array of N antenna elements spaced at half wavelength is placed horizontally at the AP with discrete Fourier transform (DFT) beamforming adopted to form N fixed beams. Without loss of generality, by assuming a line-of-sight (LoS) channel at mmWave frequencies, the beam gain of beam n numbered from the left-hand side to the right-hand side naturally for user u_k is given by $D_{k,n} = \frac{\sin^2\left(\frac{N\pi}{2}\cos\theta_k - \psi_n\right)}{N\sin^2\left(\frac{\pi}{2}\cos\theta_k - \frac{1}{N}\psi_n\right)}$, where θ_k denotes the angle of departure (AoD) of the signal received at user u_k, and $\psi_n = \left(-\frac{N+1}{2} + n\right)\pi$ [47].

1) Decomposition Results: Fig. 9 illustrates the decomposition results of a fixed beam based single-cell massive MIMO system with the proposed RC-NetDecomp algorithm when the per-user rate constraint R_{th} is 5 bit/s/Hz, 4 bit/s/Hz, 3 bit/s/Hz and 0. Similar to the cell-free massive MIMO case, as the per-user rate constraint R_{th} increases, more and more beams are grouped together to serve users jointly for the sake of inter-beam interference cancellation. In the scenario with very low or even no rate requirement, a beam is associated with the users falling into its beam coverage as depicted in Fig. 9(d), reducing to the special case of beam allocation based massive MIMO system investigated in [47], where a user is served by a single beam independently. In addition, it can be seen that with the proposed RC-NetDecomp algorithm, there are beams associated with no users, which can be turned off to save not only transmit power but also the energy consumed by radio frequency (RF) chains, as the number of RF chains in use is now smaller than the total number of beams N.

Fig. 10 presents the average number of subnetworks $\bar{M}^*$ together with the average maxi-
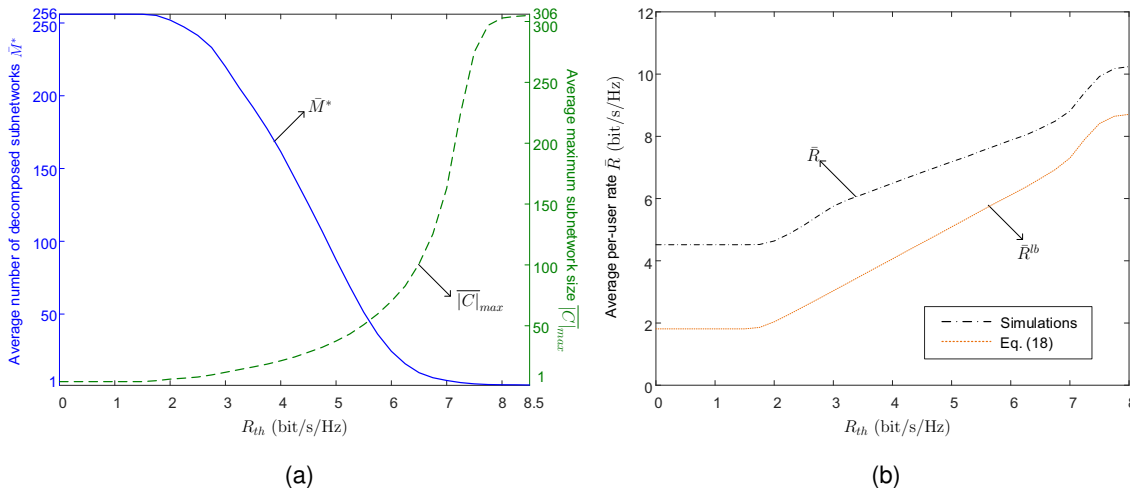
Fig. 10. (a) Average number of decomposed subnetworks $\bar{M}^*$ and average maximum subnetwork size $\overline{|C|}_{max}$, and (b) average per-user rate $\bar{R}$ versus the per-user rate constraint R_{th} with the proposed RC-NetDecomp algorithm. $N = 256$, $K = 50$, $P_t/\sigma^2 = 0$dB, $\alpha = 2.7$.

mum subnetwork size $\overline{|C|}_{max}$, and the average per-user rate $\bar{R}$ by adopting the proposed RC-NetDecomp algorithm. The average per-user rate lower-bound $\bar{R}^{lb}$ is also plotted in Fig. 10(b) for comparison. Simulation results corroborate that the proposed RC-NetDecomp algorithm can always meet the per-user rate requirement. Similar to the special case of cell-free massive MIMO, we can see from Fig. 10(a) that the rate constraint R_{th} could be carefully tuned to balance the per-user rate performance and the system complexity/signaling overhead in a fine-grained manner.

2) Comparison with User-Centric Clustering: By applying the user-centric clustering algorithm in [14] in the single-cell massive MIMO system, each user first selects its S best beams with the highest beam gains and then group the users with overlapped associated beams, which is the same as the first step of the beam-user grouping algorithm proposed in [42]. Fig. 11 presents the decomposition results with the user-centric clustering algorithm in [14] under the same user locations shown in Fig. 9. By comparing Fig. 9 and Fig. 11, it can be found that the proposed RC-NetDecomp algorithm enables much more flexible user-beam association than the user-centric clustering algorithm. For instance, in Fig. 9(c), a user close to the main direction of its best beam is served by this beam only, while a user located at the angular edge of two adjacent beams is served by these two beams jointly. By contrast, with the user-centric clustering algorithm, the number of beams chosen by each user is a constant, which might lead to unnecessary beams being activated. As illustrated in Fig. 11(b), for example, a user close to the main direction of
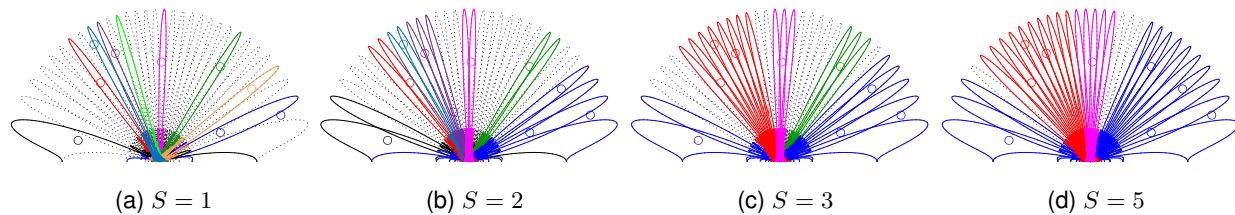
(a) $S = 1$ (b) $S = 2$ (c) $S = 3$ (d) $S = 5$

Fig. 11. Decomposition results of a fixed-beam based single-cell massive MIMO system with the user-centric clustering algorithm in [14]. "∘" represents a user. Users and beams in the same color belong to the same subnetwork. Inactive beams are drawn in dashed lines. $N = 32$, $K = 10$.
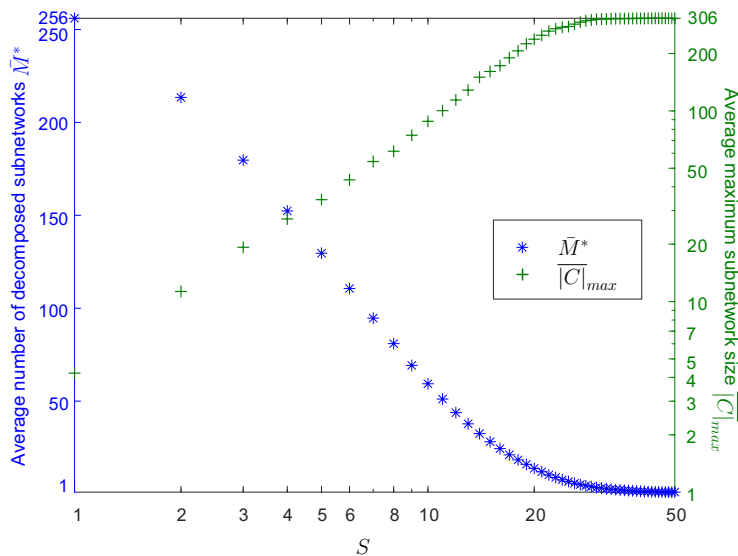

Fig. 12. Average number of decomposed subnetworks $\bar{M}^*$ and average maximum subnetwork size $\overline{|C|}_{max}$ with the user-centric clustering algorithm in [14]. $N = 256$, $K = 50$.

a beam is still served by two beams, which is unnecessary.

Fig. 12 presents the average number of decomposed subnetworks $\bar{M}^*$ and the average maximum subnetwork size $\overline{|C|}_{max}$ by varying the number of beams chosen by each user S. It can be clearly seen in this figure that the drawbacks of the user-centric approach identified in the cell-free massive MIMO special case remain here in the single-cell massive MIMO scenario. Specifically, as shown in Fig. 12, the average maximum subnetwork size $\overline{|C|}_{max}$ increases dramatically with S, implying high joint processing complexity and signaling overhead. More importantly, there are only a small number of options to adjust the number of decomposed subnetworks and the subnetwork size, which limits the practical use of the user-centric clustering benchmark for
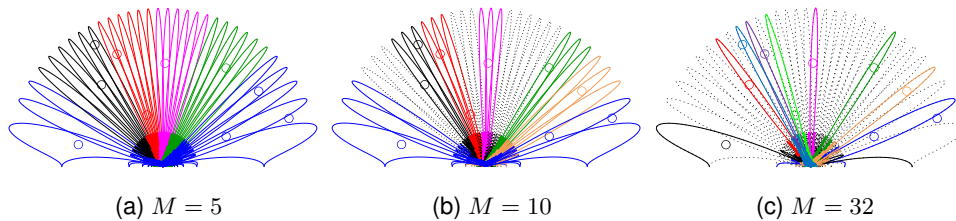
(a) $M = 5$ (b) $M = 10$ (c) $M = 32$

Fig. 13. Decomposition results of a fixed-beam based single-cell massive MIMO system with the AP-centric algorithm in [34] under the user locations presented in Fig. 9. "∘" represents a user. Users and beams in the same color belong to the same subnetwork. Inactive beams are drawn in dashed lines. $N = 32$, $K = 10$.

clustered cell-free networking.

3) Comparison with AP-Centric Clustering: To apply the AP-centric clustering algorithm in [34] in a single-cell massive MIMO system, distances between beams need to be defined. As inter-beam interference decreases with the angular separation between two beams, and the first beam and the last beam are adjacent, the distance between any beam b_i and beam b_j can be set as $\min\{\cos\theta_i - \cos\theta_j, \cos\theta_j - \cos\theta_i + 2\}$ with $\theta_i \leq \theta_j$, where θ_i is the main direction of beam b_i. Fig. 13 presents the decomposition results with the AP-centric clustering algorithm in [34] by varying the number of decomposed subnetworks M. It can be seen that different subnetworks contain a similar amount of beams, which limits the flexibility to group beams and users.

Figs. 14(a)–14(c) present the average per-user rate $\bar{R}$, the average minimum rate of users $\bar{R}_{min}$ and the average variance of users' rates $\bar{R}_{var}$ versus the number of decomposed subnetworks M, and Fig. 14(d) presents the average percentage of beams switched off $\bar{P}_{off}$ versus the average per-user rate $\bar{R}$ with both the proposed RC-NetDecomp algorithm and the AP-centric clustering algorithm in [34]. It can be seen that the proposed RC-NetDecomp algorithm outperforms the AP-centric clustering algorithm in terms of average per-user rate, user fairness by achieving a higher average minimum rate of users $\bar{R}_{min}$ and a lower variance of users' rates $\bar{R}_{var}$, and energy efficiency by switching off more beams while achieving the same average per-user rate. Note that in Fig. 14(c), fluctuations are observed in the average variance of users' rates with the AP-centric algorithm in [34]. This is because the algorithm in [34] was originally proposed for cell-free massive MIMO systems, where the distance measure for clustering is Euclidean distance. Here, the Euclidean distances between APs are replaced with the angular separations between beams to implement it in single-cell massive MIMO systems, whereas the beam gain from a beam to a user fluctuates as their angular separation increases.
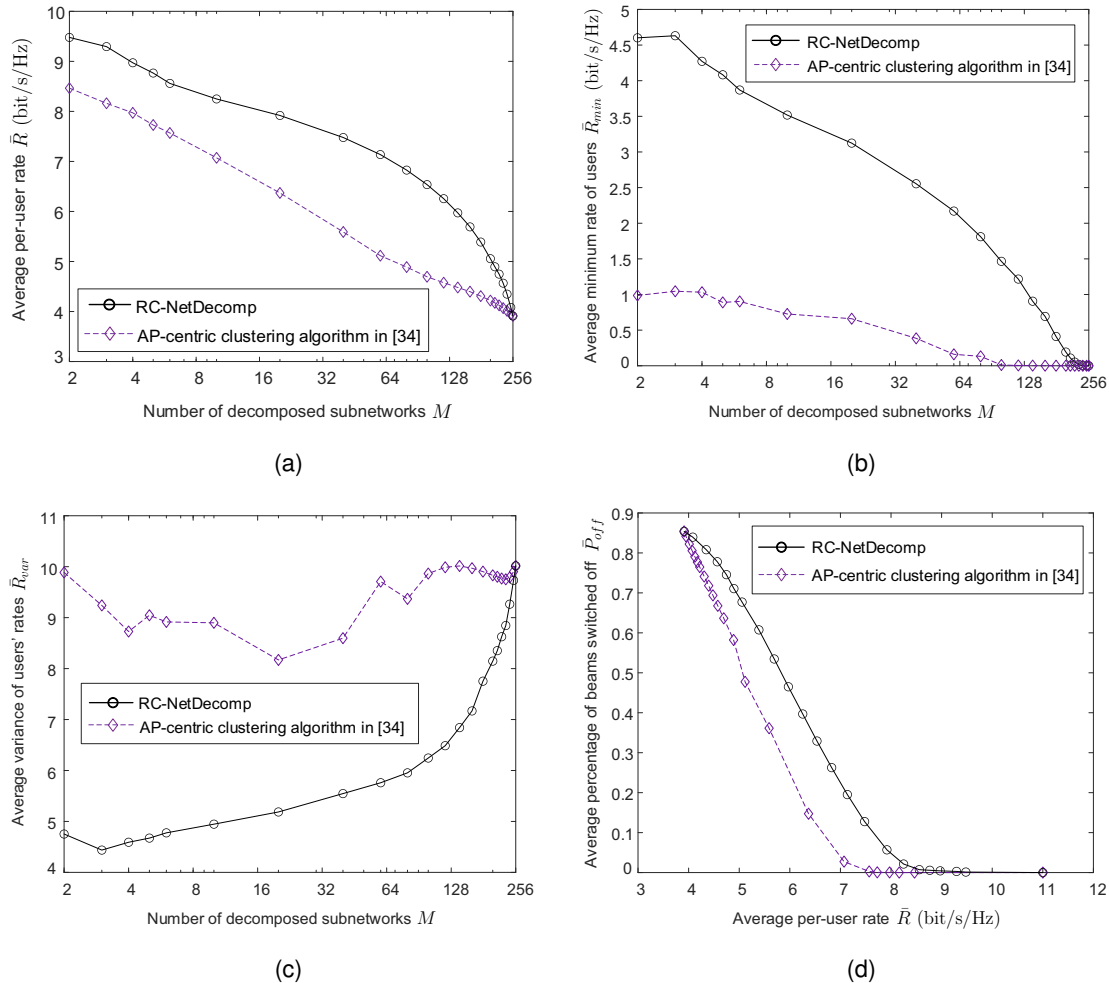
(a)

(b)

(c)

(d)

Fig. 14. (a) Average per-user rate $\bar{R}$, (b) average minimum rate of users $\bar{R}_{min}$, (c) average variance of users' rates $\bar{R}_{var}$ and (d) average percentage of beams switched off $\bar{P}_{off}$ with the proposed RC-NetDecomp algorithm and the AP-centric clustering algorithm in [34]. $N = 256$, $K = 50$, $P_t/\sigma^2 = 0$dB, $\alpha = 2.7$.

V. CONCLUSION

This paper focused on the rate-constrained clustered cell-free networking problem in a wireless network with multiple APs, where multiple beams are available at each AP. With the aim of maximizing the number of decomposed subnetworks under a per-user rate constraint, the clustered cell-free networking problem was formulated as a bipartite graph partitioning problem and a RC-NetDecomp algorithm was proposed. Since the edge weight between a user and a beam was defined as the corresponding normalized channel gain, the proposed RC-NetDecomp algorithm can switch off the beams that are badly aligned with users or associated with APs far from users to save energy, and produce subnetworks with balanced sizes at the same time. The

performance of the proposed RC-NetDecomp algorithm was evaluated and compared against the user-centric and AP-centric benchmarks in two special cases of cell-free massive MIMO and single-cell massive MIMO. Simulation results showed that in both cases, the proposed RC-NetDecomp algorithm provides a fine-grained tuning of subnetwork size and produces more balanced subnetworks with smaller maximum subnetwork size than the user-centric baseline, which can efficiently reduce the joint processing complexity and signaling overhead in practice. Compared to the AP-centric baseline, higher average per-user rate, better user fairness and higher energy efficiency can be achieved by adopting the proposed RC-NetDecomp algorithm.

Note that to implement the RC-NetDecomp algorithm, channel gains between beams and users in the network need to be collected. To reduce the channel measurement overhead, it is of paramount importance to further develop distributed clustered cell-free networking schemes in the future, where each AP only needs to measure the channels of the users within its vicinity and exchange information with a small number of nearby APs.

APPENDIX A

DERIVATION OF (30)–(32)

Recall that a new graph $\tilde{G} = (\tilde{\mathcal{V}}, \tilde{E})$ is constructed by merging each beam vertex b_i with the users vertices with edge weight 1 between them on graph G to form the meganode $\tilde{v}_i$ given in (28). Suppose that graph $\tilde{G}$ is partitioned into M subgraphs and the set of meganodes in the mth subgraph is denoted as $\tilde{\mathcal{C}}_m$ with $\bigcup_{m=1}^{M} \tilde{\mathcal{C}}_m = \tilde{\mathcal{V}}$ and $\tilde{\mathcal{C}}_m \cap \tilde{\mathcal{C}}_{m'} = \emptyset, \forall m' \neq m$. The corresponding set of user and beam vertices in the mth subgraph on graph G, C_m, is then given by

$$C_m = \bigcup_{\tilde{v}_i \in \tilde{\mathcal{C}}_m} \tilde{v}_i = \left\{ b_i : \tilde{v}_i \in \tilde{\mathcal{C}}_m \right\} \cup \left\{ u_k : u_k \in \tilde{v}_i, \tilde{v}_i \in \tilde{\mathcal{C}}_m \right\}. \tag{A.1}$$

By combining (15), (28) and (A.1), the cut function $\text{cut}(C_m)$ can be rewritten as

$$\text{cut}(C_m) = \sum_{u_k \in C_m} \sum_{b_n \in B, b_n \notin C_m} w_{k,n} + \sum_{b_n \in C_m} \sum_{u_k \in U, u_k \notin C_m} w_{k,n}$$

$$= \sum_{\tilde{v}_i \in \tilde{\mathcal{C}}_m} \sum_{u_k \in \tilde{v}_i} \sum_{\tilde{v}_j \in \tilde{\mathcal{V}}, \tilde{v}_j \notin \tilde{\mathcal{C}}_m} w_{k,j} + \sum_{\tilde{v}_i \in \tilde{\mathcal{C}}_m} \sum_{\tilde{v}_j \in \tilde{\mathcal{V}}, \tilde{v}_j \notin \tilde{\mathcal{C}}_m} \sum_{u_k \in \tilde{v}_j} w_{k,i} = \sum_{\tilde{v}_i \in \tilde{\mathcal{C}}_m} \sum_{\tilde{v}_j \in \tilde{\mathcal{V}}, \tilde{v}_j \notin \tilde{\mathcal{C}}_m} \left(\sum_{u_k \in \tilde{v}_i} w_{k,j} + \sum_{u_k \in \tilde{v}_j} w_{k,i} \right).$$

$$\tag{A.2}$$

By substituting (29) into (A.2), we have

$$\text{cut}(C_m) = \sum_{\tilde{v}_i \in \tilde{\mathcal{C}}_m} \sum_{\tilde{v}_j \in \tilde{\mathcal{V}}, \tilde{v}_j \notin \tilde{\mathcal{C}}_m} \tilde{w}_{i,j} = \text{cut}(\tilde{\mathcal{C}}_m). \tag{A.3}$$

As a meganode on the new graph $\tilde{G}$ always contains one beam, constraint (27) is always satisfied. According to (A.3), the equivalent graph partitioning problem on graph $\tilde{G} = (\tilde{\mathcal{V}}, \tilde{E})$ given in (30)–(32) can be therefore obtained from (24)–(27).

APPENDIX B

PROOF OF THEOREM 1

Proof: The mincut function $\sum_{m=1}^{M+1} \text{cut}(C_{m|M+1}^*)$ for a given number of subgraphs $M+1$ can be written as

$$\sum_{m=1}^{M+1} \text{cut}(C_{m|M+1}^*) = \sum_{m=1}^{M-1} \text{cut}(C_{m|M+1}^*) + \text{cut}(C_{M|M+1}^*) + \text{cut}(C_{M+1|M+1}^*). \qquad \text{(B.1)}$$

Recall that the cut function of a set X, $\text{cut}(X)$, is given in (14) as $\text{cut}(X) = \sum_{v_i \in X} \sum_{v_j \in \overline{X}} a_{i,j}$. For any two sets $X, Y \subseteq V$ and $X \cap Y = \emptyset$, the cut function, $\text{cut}(X \cup Y)$, can be obtained as

$$\text{cut}(X \cup Y) = \sum_{v_i \in X \cup Y} \sum_{v_j \in \overline{X \cup Y}} a_{i,j} = \sum_{v_i \in X} \sum_{v_j \in \overline{X \cup Y}} a_{i,j} + \sum_{v_i \in Y} \sum_{v_j \in \overline{X \cup Y}} a_{i,j}. \qquad \text{(B.2)}$$

For any set X, it is obvious that $\overline{X \cup Y} \subseteq \overline{X}$. We then have

$$\text{cut}(X \cup Y) \leq \sum_{v_i \in X} \sum_{v_j \in \overline{X}} a_{i,j} + \sum_{v_i \in Y} \sum_{v_j \in \overline{Y}} a_{i,j} = \text{cut}(X) + \text{cut}(Y). \qquad \text{(B.3)}$$

The mincut function $\sum_{m=1}^{M+1} \text{cut}(C_{m|M+1}^*)$ given in (B.1) is then lower-bounded by

$$\sum_{m=1}^{M+1} \text{cut}(C_{m|M+1}^*) \geq \sum_{m=1}^{M-1} \text{cut}(C_{m|M+1}^*) + \text{cut}(C_{M|M+1}^* \cup C_{M+1|M+1}^*) \geq \sum_{m=1}^{M} \text{cut}(C_{m|M}^*), \quad \text{(B.4)}$$

as $\mathcal{M}_{|M}^* = \left\{ C_{1|M}^*, C_{2|M}^*, \cdots C_{M|M}^* \right\}$ is the optimal M-way partition that minimizes the cut function $\sum_{m=1}^{M} \text{cut}(C_m)$. ∎

REFERENCES

[1] J. Wang, L. Dai, L. Yang, and B. Bai, "Rate-constrained network decomposition for clustered cell-free networking," to appear in *IEEE ICC*, May 2022.

[2] W. Saad, M. Bennis, and M. Chen, "A vision of 6G wireless systems: Applications, trends, technologies, and open research problems," *IEEE Network*, vol. 34, no. 3, pp. 134–142, May/June 2020.

[3] H. Zhuang, L. Dai, L. Xiao, and Y. Yao, "Spectral efficiency of distributed antenna system with random antenna layout," *Electronics Letters*, vol. 39, no. 6, pp. 495–496, Mar. 2003.

[4] J. Zhang and J. G. Andrews, "Distributed antenna systems with randomness," *IEEE Trans. Wireless Commun.*, vol. 7, no. 9, pp. 3636–3646, Sep. 2008.

[5] H. Hu, Y. Zhang, and J. Luo, *Distributed Antenna Systems: Open Architecture for Future Wireless Communications*, CRC Press, 2007.

[6] L. Dai, "A comparative study on uplink sum capacity with co-located and distributed antennas," *IEEE J. Select. Areas Commun.*, vol. 29, no. 6, pp. 1200–1213, June 2011.

[7] Z. Liu and L. Dai, "A comparative study of downlink MIMO cellular networks with co-located and distributed base-station antennas," *IEEE Trans. Wireless Commun.*, vol. 13, no. 11, pp. 6259–6274, Nov. 2014.

[8] J. Wang and L. Dai, "Asymptotic rate analysis of downlink multi-user systems with co-located and distributed antennas," *IEEE Trans. Wireless Commun.*, vol. 14, no. 6, pp. 3046–3058, June 2015.

[9] China Mobile, "C-RAN: The road toward green RAN," White Paper, ver. 2.5, Oct. 2011.

[10] H. Q. Ngo, A. Ashikhmin, H. Yang, E. G. Larsson, and T. L. Marzetta, "Cell-free massive MIMO versus small cells," *IEEE Trans. Wireless Commun.*, vol. 16, no. 3, pp. 1834–1850, Mar. 2017.

[11] L. Dai, S. Zhou, and Y. Yao, "Capacity with MRC-based macrodiversity in CDMA distributed antenna systems," in *Proc. IEEE Globecom*, pp. 987–991, Nov. 2002.

[12] L. Dai, S. Zhou, and Y. Yao, "Capacity analysis in CDMA distributed antenna systems," *IEEE Trans. Wireless Commun.*, vol. 4, no. 6, pp. 2613–2620, Nov. 2005.

[13] L. Dai, "An uplink capacity analysis of the distributed antenna system (DAS): From cellular DAS to DAS with virtual cells," *IEEE Trans. Wireless Commun.*, vol. 13, no. 5, pp. 2717-–2731, May 2014.

[14] J. Wang and L. Dai, "Downlink rate analysis for virtual-cell based large-scale distributed antenna systems," *IEEE Trans. Wireless Commun.*, vol. 15, no. 3, pp. 1998–2011, Mar. 2016.

[15] T. C. Mail, H. Q. Ngo, M. Egan, and T. Q. Duong, "Pilot power control for cell-free massive MIMO," *IEEE Trans. Veh. Tech.*, vol. 67, no. 11, pp. 11264–11268, Nov. 2018.

[16] M. Attarifar, A. Abbasfar, and A. Lozano, "Subset MMSE receivers for cell-free networks," *IEEE Trans. Wireless Commun.*, vol. 19, no. 6, pp. 4183–4194, June 2020.

[17] R. Mosayebi, M. M. Mojahedian, and A. Lozano, "Linear interference cancellation for the cell-free C-RAN uplink," *IEEE Trans. Wireless Commun.*, vol. 20, no. 3, pp. 1544–1556, Mar. 2021.

[18] S. Buzzi and C. D'Andrea, "Cell-free massive MIMO: User-centric approach," *IEEE Wireless Commun. Lett.*, vol. 6, no. 6, pp. 706–709, Dec. 2017.

[19] M. Alonzo, S. Buzzi, A. Zappone, and C. D'Elia, "Energy-efficient power control in cell-free and user-centric massive MIMO at millimeter wave," *IEEE Trans. Green Commun. Net.*, vol. 3, no. 3, pp. 651–663, Sept. 2019.

[20] Y. Lin, R. Zhang, C. Li, L. Yang, and L. Hanzo, "Graph-based joint user-centric overlapped clustering and resource allocation in ultradense networks," *IEEE Trans. Veh. Tech.*, vol. 67, no. 5, pp. 4440–4453, May 2018.

[21] Y. Lin, R. Zhang, L. Yang, and L. Hanzo, "User-centric clustering for designing ultradense networks: Architecture, objective functions and design guidelines," *IEEE Veh. Tech. Mag.*, vol. 14, no. 3, pp. 107–114, Sept. 2019.

[22] S. Mukherjee and J. Lee, "Edge computing-enabled cell-free massive MIMO systems," *IEEE Trans. Wireless Commun.*, vol. 19, no. 4, pp. 2884–2899, Apr. 2020.

[23] G. Interdonato, M. Karlsson, E. Bjornson, and E. G. Larsson, "Local partial zero-forcing precoding for cell-free massive MIMO," *IEEE Trans. Wireless Commun.*, vol. 19, no. 7, pp. 4758–4774, July 2020.

[24] H. Liu, J. Zhang, S. Jin, and B. Ai, "Graph coloring based pilot assignment for cell-free massive MIMO systems," *IEEE Trans. Veh. Tech.*, vol. 69, no. 8, pp. 9180–9184, Aug. 2020.

[25] L. Dai and B. Bai, "Optimal decomposition for large-scale infrastructure-based wireless networks," *IEEE Trans. Wireless Commun.*, vol. 16, no. 8, pp. 4956–4969, Aug. 2017.

[26] M. K. Karakayali, G. J. Foschini, and R. A. Valenzuela, "Network coordination for spectrally efficient communications in cellular systems," *IEEE Wireless Commun.*, vol. 13, pp. 56–61, Aug. 2006.

[27] J. Zhang, R. Chen, J. G. Andrews, A. Ghosh, and R. W. Heath, Jr., "Network MIMO with clustered linear precoding," *IEEE Trans. Wireless Commun.*, vol. 8, no. 4, pp. 1910–1921, Apr. 2009.

[28] H. Dahrouj and W. Yu, "Coordinated beamforming for the multicell multi-antenna wireless system," *IEEE Trans. Wireless Commun.*, vol. 9, no. 5, pp. 1748–1759, May 2010.

[29] K. Huang and J. G. Andrews, "An analytical framework for multicell cooperation via stochastic geometry and large deviations," *IEEE Trans. Inf. Theory*, vol. 59, no. 4, pp. 2501–2516, Apr. 2013.

[30] A. Papadogiannis, D. Gesbert, and E. Hardouin, "A dynamic clustering approach in wireless networks with multi-cell cooperative processing," in *Proc. IEEE ICC*, pp. 4033–4037, May 2008.

[31] W. Saad, Z. Han, M. Debbah, and A. Hjorungnes, "A distributed coalition formation framework for fair user cooperation in wireless networks," *IEEE Trans. Wireless Commun.*, vol. 8, no. 9, pp. 4580–4593, Sept. 2009.

[32] N. Lee, D. Morales-Jimenez, A. Lozano, and R. W. Heath, "Spectral efficiency of dynamic coordinated beamforming: A stochastic geometry approach," *IEEE Trans. Wireless Commun.*, vol. 14, no. 1, pp. 230–241, Jan. 2015.

[33] M. Karlksson, E. Bjornson, and E. G. Larsson, "Techniques for system information broadcast in cell-free massive MIMO," *IEEE Trans. Commun.*, vol. 67, no. 1, pp. 244–257, Jan. 2019.

[34] M. Yemini and A. J. Goldsmith, "Virtual cell clustering with optimal resource allocation to maximize cellular system capacity," in *Proc. IEEE Globecom*, Dec. 2019.

[35] J. Bien, and R. Tibshirani, "Hierarchical clustering with prototypes via minimax linkage," *J. Am. Stat. Assoc.*, vol. 106, no. 495, pp. 1075–1084, 2011.

[36] C. Xiao, J. Zeng, X. Su, J. Wang, X. Xu, L. Ge, and L. Zhang, "Downlink transmission scheme based on virtual cell merging in ultra dense networks," in *Proc. VTC-Fall*, Sept. 2016.

[37] L. Chen, H. Xia, Z. Zeng, and S. Wu, "Joint merging overlapped virtual cells and designing beamforming in ultra dense network," in *Proc. IEEE ISWCS*, Aug. 2017.

[38] Y. Lin, R. Zhang, L. Yang, and L. Hanzo, "Modularity-based user-centric clustering and resource allocation for ultra dense networks," *IEEE Trans. Veh. Tech.*, vol. 67, no. 12, pp. 12457–12461, Dec. 2018.

[39] G. Femenias and F. Riera-Palou, "Cell-free millimeter-wave massive MIMO systems with limited fronthaul capacity," *IEEE Access*, vol. 7, pp. 44596–44612, Apr. 2019.

[40] X. Zhang, J. Wang, and H. V. Poor, "Statistical delay and error-rate bounded QoS provisioning over mmWave cell-free M-MIMO and FBC-HARQ-IR based 6G wireless networks," *IEEE J. Sel. Areas Commun.*, vol. 38, no. 8, pp. 1661–1677, Aug. 2020.

[41] A. Adhikary, J. Nam, J.-Y. Ahn, G. Caire, "Joint spatial division and multiplexing–The large-scale array regime," *IEEE Trans. Inf. Theory*, vol. 59, no. 10, pp. 6441–6463, Oct. 2013.

[42] J.-S. Sheu, W.-H. Sheen, and T.-X. Guo, "On the design of downlink multiuser transmission for a beam-group division 5G system," *IEEE Trans. Veh. Tech.*, vol. 67, no. 8, pp. 7191–7203, Aug. 2018.

[43] S. T. Veetil, K. Kuchi, and R. K. Ganti, "Coverage analysis of cloud radio networks with finite clustering," *IEEE Trans. Wireless Commun.*, vol. 16, no. 1, pp. 594–606, Jan. 2017.

[44] Z. Liu,"On the scaling behavior of the average rate performance of large-scale distributed MIMO systems," *IEEE Trans. Veh. Tech.*, vol. 66, no. 5, May 2017.

[45] J. Wang, H. Zhu, Y. Kai, "On the performance of beam allocation based multi-user massive MIMO systems," in *Proc. IEEE ICC*, May 2019.

[46] U. von Luxburg, "A tutorial on spectral clustering," *Statistics and Computing*, vol. 17, no. 4, pp. 395–416, Dec. 2007.

[47] J. Wang, H. Zhu, L. Dai, N. J. Gomes, and J. Wang, "Low-complexity beam allocation for switched-beam based multiuser massive MIMO systems," *IEEE Trans. Wireless Commun.*, vol. 15, no. 12, pp. 8236–8248, Dec. 2016.