Graph-Based Clustering and Pilot Assignment for Scalable Cell-Free Massive MIMO Systems

Bariq S. Firmansyah and Irma Zakia
School of Electrical Engineering and Informatics, Institut Teknologi Bandung, Indonesia

***Abstract*—Cell-Free Massive MIMO is a promising technology which many access points cooperate to jointly serve UEs within the same time-frequency resource. However, the original form of Cell-Free Massive MIMO is not scalable when the number of UEs goes to infinity. Clustering is one of the solutions to make Cell-Free Massive MIMO systems becomes scalable where each AP only serves a subset of UEs with the best channels. In this paper, we propose a framework to form clusters between UE and APs and pilot assignment that takes the maximum traffic load of each AP into consideration. The proposed algorithm provides 54% improvement in 95%-likely per-user SE over the existing scheme when the number of UEs is massive.**

***Index Terms*—Cell-Free Massive MIMO, Clustering, Gale-Shapley, Graph coloring, Beyond 5G**

I. Introduction

Cell-Free Massive MIMO was first introduced in [1] in which a massive number of distributed access points cooperate to jointly serve smaller number of user equipments in the same time-frequency resources. The APs are connected via fronthaul network to central processing unit (CPU) which sends the downlink data symbols to all APs and decode the uplink data either in distributed fashion where each AP sends the soft-estimate of the received data symbols or centralized fashion where APs forwards all the received uplink pilot and data symbols to CPU [2]. Cell-Free Massive MIMO provides a uniform performance among the users [3] in contrast to the conventional cellular networks where the edge-users perform very poor due to significant path-loss and strong inter-cell interference from the neighboring cells.

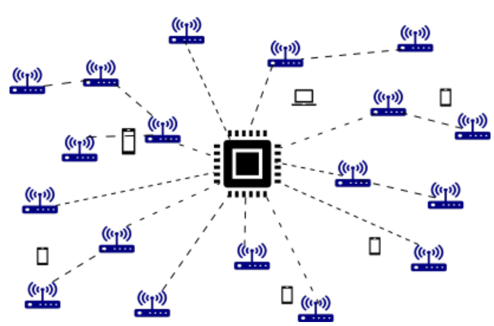

Fig. 1. Cell-Free Massive MIMO Network

User-centric approach of cell-free massive MIMO was introduced in [4] where the AP does not serve all UEs in the network, but only the UEs with the strongest channels. In [4] [5], it was shown that the user-centric approach significantly reduced the fronthaul complexity while maintaining a good system performance. The scalability aspect of cell-free massive MIMO was studied in [6]. It was shown that the original form of cell-free massive MIMO is not scalable because the computational complexity goes to infinity when the number of UEs grows very big.

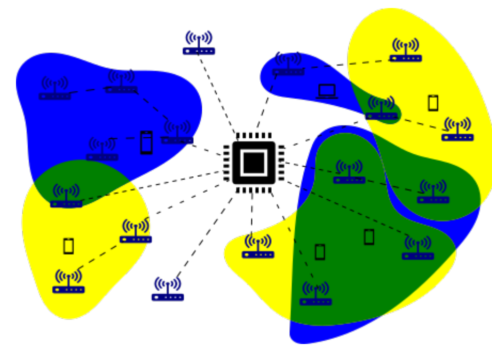

Fig. 2. User-centric Cell-Free Massive MIMO

The algorithm for clustering and pilot assignment were also proposed in [6] [7] with the assumption that the maximum number of UEs served by an AP equals the number of orthogonal pilot (i.e., $|\mathcal{D}_l| \leqslant \tau_p$). In this paper, we propose a framework for clustering and pilot assignment which consider the traffic limit of an AP as the maximum number of UEs served by an AP (i.e., $|\mathcal{D}_l| \leqslant S_l$). The proposed framework outperforms the benchmark in terms of 95%-likely per-user SE when the number of UE is massive, i.e, $K \approx L$.

Paper Organization: The paper is organized as follows. System Model is described in Section II. In Section III, we explain the proposed clustering framework. Section IV discuss the numerical results. Section V concludes the paper.

II. System Model

The system consists of L APs equipped with N antennas and K single-antenna UEs which are randomly distributed in large area. The L APs are connected to CPU via fronthaul link as illustrated in Fig. 2. We only consider the uplink transmissions where all the communications take place on the same frequency band. TDD operation is used in the communication between APs and UEs where each coherence block τ_c is divided into τ_p channels used for uplink pilot training and $\tau_u = \tau_c - \tau_p$ channels used for uplink data. The channel between AP l and UE k is denoted by $\mathrm{h}_{kl} \in \mathbb{C}^N$ for $k = 1, ..., K$ and $l = 1, ..., L$. In each coherence block, an independent channel realization from correlated rayleigh fading distribution is drawn by $\mathrm{h}_{kl} \sim \mathcal{N}_c(0, \mathrm{R}_{kl})$ where $\mathrm{R}_{kl} \in \mathbb{C}^{N\times N}$ is the spatial correlation matrix that describes the large-scale fading property of the channel. The connection between AP l and UE k are denoted by diagonal matrix $\mathrm{D}_{kl} \in \mathbb{C}^{N\times N}$ in which the nth diagonal value indicates whether antenna n of AP l is allowed to decode and transmit to UE k; 1 is allowed, 0 otherwise. Moreover, matrix $\mathrm{A} \in \mathbb{R}^{L\times K}$ is defined to specify the clustering between UEs and APs where the entry $\mathrm{A}_{kl} = 1$ if UE k is served by at least one antenna of AP l (i.e., $\mathrm{tr}(\mathrm{D}_{kl}) > 0$) and 0 otherwise. The subset of APs serving UE k is denoted by $\mathcal{D}_l = \{k : \mathrm{A} = 1, k \in \{1, ..., K\}\}$.

A. Pilot Training and Channel Estimation

We assume there are τ_p mutually orthogonal τ_p-length pilot signals $\phi_1, ..., \phi_{\tau_p}$ satisfying $\|\phi_t\|^2 = \tau_p$. The number of τ_p depends on the graph coloring-based pilot assignment output and of course we must ensure that $\tau_p < \tau_c$. We consider a scenario where the number of UEs is larger than the number of pilot signals (i.e., $K > \tau_p$) so that there are UEs that share the same pilot, these UEs are referred to as pilot-sharing UEs. We denote $t_k \in \{1, ..., \tau_p\}$ as the index of the pilot assigned to UE k. $\mathcal{T}_k$ is defined as the set of UEs assigned to pilot signal ϕ_{t_k}. AP l received the pilot signal $\mathrm{y}^{\mathrm{pilot}}_{t_k l} \in \mathbb{C}^N$ from UEs in $\mathcal{T}_k$ [8], and is given by.

$$\mathrm{y}^{\mathrm{pilot}}_{t_k l} = \sum_{i\in\mathcal{T}_k} \sqrt{\tau_p p_i}\mathrm{h}_{il} + \mathrm{n}_l \tag{1}$$

where p_i denotes the transmit power of UE i, τ_p denotes the processing gain, and $\mathrm{n}_l \sim \mathcal{N}_{\mathbb{C}}(0, \sigma^2\mathrm{I}_N)$ is thermal noise. Channel between UE $k \in \mathcal{T}_k$ and AP l h_{kl} is estimated using minimum mean-squared-error (MMSE) estimation [8], and is given by

$$\hat{\mathrm{h}}_{kl} = \sqrt{\tau_p p_k}\mathrm{R}_{kl}\boldsymbol{\psi}^{-1}_{t_k l}\mathrm{y}^{\mathrm{pilot}}_{t_k l} \tag{2}$$

where

$$\boldsymbol{\psi}_{t_k l} = \mathbb{E}\left\{\mathrm{y}^{\mathrm{pilot}}_{t_k l}(\mathrm{y}^{\mathrm{pilot}}_{t_k l})^{\mathrm{H}}\right\} = \sum_{i\in\mathcal{T}_k} \tau_p p_i \mathrm{R}_{il} + \sigma^2\mathrm{I}_N \tag{3}$$

denotes the correlation matrix of (1). Channel estimate $\hat{\mathrm{h}}_{kl}$ and estimation error $\tilde{\mathrm{h}}_{kl}$ are Gaussian distributed with 0 mean and B_{kl} and C_{kl} variance respectively, where

$$\mathrm{B}_{kl} = \mathbb{E}\left\{\hat{\mathrm{h}}_{kl}\hat{\mathrm{h}}^{\mathrm{H}}_{kl}\right\} = \tau_p p_k \mathrm{R}_{kl}\boldsymbol{\psi}^{-1}_{t_k l}\mathrm{R}_{kl} \tag{4}$$

is the correlation matrix of (2) and

$$\mathrm{C}_{kl} = \mathbb{E}\left\{\tilde{\mathrm{h}}_{kl}\tilde{\mathrm{h}}^{\mathrm{H}}_{kl}\right\} = \mathrm{R}_{kl} - \mathrm{B}_{kl} \tag{5}$$

is the correlation matrix of $\tilde{\mathrm{h}}_{kl}$

Eq. (1) shows us that pilot-sharing UEs $\mathcal{T}_k$ degrades the quality of estimated channel $\hat{\mathrm{h}}_{kl}$ in a sense that the estimates $\hat{\mathrm{h}}_{kl}$ for $k \in \mathcal{T}_k$ become correlated. This phenomenon is known as pilot contamination. In this paper, we consider the distributed processing scheme where each AP estimates the channel of its served UEs in $\mathcal{D}_l$.

B. Uplink Data Transmission

In the uplink data transmission, every AP receives transmitted symbols from all UEs in the network, the received signal $\mathrm{y}^{\mathrm{ul}}_l \in \mathbb{C}^N$ is expressed as

$$\mathrm{y}^{\mathrm{ul}}_l = \sum_{i=1}^{K} \mathrm{h}_{il}s_i + \mathrm{n}_l \tag{6}$$

where $s_i \in \mathbb{C}$ is a complex symbol transmitted from UE i with power p_i and $\mathrm{n}_l \sim \mathcal{N}_{\mathbb{C}}(0, \sigma^2\mathrm{I}_N)$ is the thermal noise in AP l receiver.

In order to avoid overloaded CPU, we consider the distributed processing scheme in which the uplink data are computed locally in every AP by selecting the local combining vector v_{kl} based on its local channel estimates $\hat{\mathrm{h}}_{kl}$ and send the local estimate

$$\hat{s}_{kl} = \mathrm{v}^{\mathrm{H}}_{kl}\mathrm{D}_{kl}\mathrm{y}^{\mathrm{ul}}_l \tag{7}$$

to a CPU to compute the final estimate of s_k

$$\hat{s}_k = \sum_{l=1}^{L} \hat{s}_{kl} \tag{8}$$

Eq. (5) shows us that any combining vector can be applied to compute the local estimate $\hat{s}_{kl}$. Note that we should select the combining vector that is proven to be scalable. Local Partial MMSE (LP-MMSE) combining, considered in [6] as the scalable alternative of MMSE combining, can be expressed as

$$\hat{\mathrm{v}}^{\mathrm{LP\text{-}MMSE}}_{kl} = p_k\Big(\sum_{i\in\mathcal{D}_l} p_i(\hat{\mathrm{h}}_{il}\hat{\mathrm{h}}^{\mathrm{H}}_{il} + \mathrm{C}_{il}) + \sigma^2\mathrm{I}_N\Big)^{-1}\hat{\mathrm{h}}_{kl} \tag{9}$$

LP-MMSE combining exploits the fact that the main source of interference on UE k comes from the other UEs that AP l is currently serving (i.e., $i \in \mathcal{D}_l, i \neq k$)

CPU does not have knowledge of channel estimate since the channel estimation is performed locally in AP. We utilize use-and-then-forget (UatF) bound that is wideluy used in Massive MIMO [8] to compute an achievable uplink SE

Proposition 1. An achievable UL SE for UE k is

$$\mathrm{SE}^{\mathrm{ul}}_k = \frac{\tau_u}{\tau_c}\log_2\left(1 + \mathrm{SINR}^{\mathrm{ul}}_k\right) \tag{10}$$

where

$$\mathrm{SINR}^{\mathrm{ul}}_k = \frac{p_k|\mathbb{E}\{\mathrm{v}^{\mathrm{H}}_k\mathrm{D}_k\mathrm{h}_k\}|^2}{\sum_{i=1}^{K} p_i\mathbb{E}\{|\mathrm{v}^{\mathrm{H}}_k\mathrm{D}_k\mathrm{h}_i|^2\} - p_k|\mathbb{E}\{\mathrm{v}^{\mathrm{H}}_k\mathrm{D}_k\mathrm{h}_k\}|^2 + \sigma^2_{\mathrm{ul}}\mathbb{E}\{\|\mathrm{D}_k\mathrm{v}_k\|\}^2} \tag{11}$$

Proof: It follows the similar approach as in [8], but for the received signal in (6). The expectation in the SINR expression for LP-MMSE combiner cannot be computed in closed-from but can be easily computed using Monte-Carlo simulations

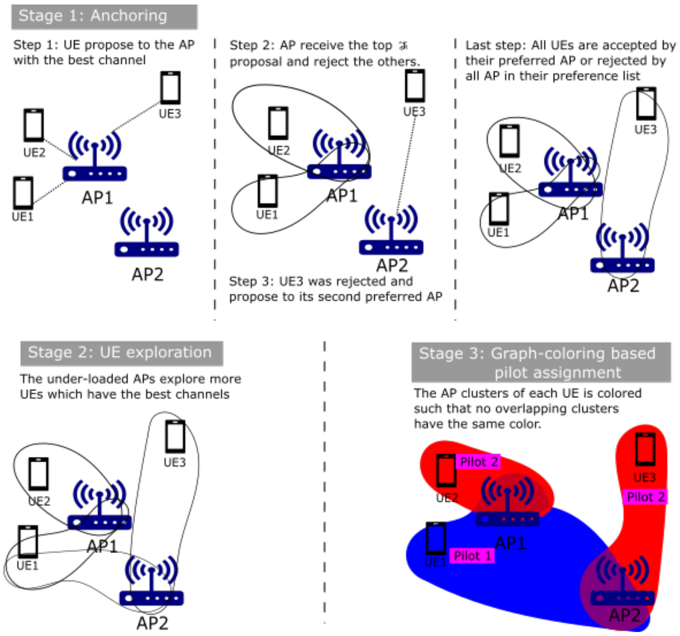

Fig. 3. The proposed procedure for clustering and pilot assignment with maximum traffic load of each AP $\mathcal{S}_l = 2$

III. Proposed Framework

There are practical constraints that needs to be considered when forming user-centric overlapped clusters, such as the maximum traffic-load of each AP. Clustering and pilot assignment frameworks for scalable cell-free massive MIMO system have been considered in [6] [7]. However, [6] didn't consider a clustering scheme in which each AP has the maximum traffic-load $\mathcal{S}_l$. In our work, we design a clustering algorithm where the number of served UEs is limited by the maximum traffic-load of each AP.

A. Clustering

In this work, we design a clustering algorithm where the number of UEs served by AP l is limited by its maximum traffic-load (i.e., $|\mathcal{D}_l| \leqslant \mathcal{S}_l$). The large-scale fading coefficient $\beta_{kl} \triangleq \text{tr}(\text{R}_{kl})/N$ is selected as the criterion to determine the relationship between APs and UEs. Inspired by [9], the user-centric clustering process consists of two stages: anchoring stage and exploration stage.

1) Anchoring stage, inspired by Gale-Shapley algorithm [10], matches each UE to its anchor AP. In this stage, each UE only connects to one AP and each AP can serve more than one UEs, thus the matching is categorized as many-to-one matching problem. Large-scale fading coefficient is selected as the preference profile of UEs and APs since it is related to the channel condition such as path loss and shadowing. Each UE receives the synchronization signals from its surrounding APs and measures the β_{kl}. The measurement of β_{kl} will determine UE's preference list. Once the preference list is formed, each UE proposes to its most preferred AP. The AP proposed by UEs can choose to reject or accept the UE proposal. If the number of proposal is larger than the maximum traffic load, the AP will keep the best $\mathcal{S}_l$ UEs and reject the rest UEs. The rejected UEs will propose to their next preferred APs until their proposal get accepted or rejected by all APs in their preference lists. Note that the rejection mechanism offered by the modified Gale-Shapley algorithm ensure the AP to serve no more than $\mathcal{S}_l$ UEs.

2) Exploration stage: After the anchoring stage, the under-loaded APs will explore more UEs. The APs will serve the best UEs which have the large-scale fading coefficient greater than an arbitrary threshold i.e., $\beta_{kl} \geqslant \delta$. The value of δ can not be lower than zero because $\beta_{kl} = 0$ implies that the path-loss is infinity. Setting δ to zero will make the AP explore all UEs in the network. Hence, the δ must be set accordingly such that it could explore around $\mathcal{S}_l$ UEs. If the number of UEs explored added by the number of UEs from anchoring stage is greater than $\mathcal{S}_l$, AP must sort the explored UEs and serve only the best UEs such that the number of served UEs is equal to $\mathcal{S}_l$. At the end of the exploration stage, the UEs may be served by more than one APs, forming user-centric AP clusters for each UE. The overall association constitutes a many-to-many matching between APs and UEs.

B. Pilot Assignment

Once the user-centric AP clusters are formed, APs inform the set of UEs they served to the CPU. The CPU will assign the pilot based on the information provided by APs using Dsatur graph-coloring algorithm [11]. Each UE represents a vertex and the edge between two vertices represents overlapping clusters which happened when two UEs k and i are served by partially the same APs, i.e., $\text{D}_k\text{D}_i \neq 0_{LN}$, where $\text{D}_k = \text{diag}(\text{D}_{k1}, ..., \text{D}_{kL}) \in \mathbb{C}^{M \times M}$ is a block-diagonal matrix. Graph coloring aims to color each vertex of a graph using the minimum number of colors possible such that no adjacent vertices get the same color. In cell-free context, the overlapping clusters will not be assigned the same pilot as it will degrade the channel estimation quality.

The problem with pilot assignment is that the number of coherence time is limited. We have to ensure that the number of colors (pilots) required to color (assign) the vertices (UEs) is less than the number of coherence time, i.e., $\tau^* < \tau_c$. In [12], if the number of required pilots is larger than the number of provided pilots, i.e., $\tau^* > \tau_p$, the author proposes to decrease the number of APs connected to each UE in order to make the required number of pilots equal to the number of provided pilots, i.e., $\tau^* = \tau_p$. However, the reduction of the number of APs serving a UE will reduce the performance of a UE. In this paper, we propose a different approach in which the APs do not decrease the number of their served UEs unless the number of required pilot exceeds the number of coherence time, i.e., $\tau^* \geqslant \tau_c$. Hence, the number of pilots depends on the number of required pilots instead of a fixed number. The number of pilot will be set based on the number of required pilots, i.e., $\tau_p = \tau^*$. The reduction of the number of served UEs can be done by increasing the threshold δ and redoing the exploration stage. The illustration and the detailed procedure are described in Figure 3.

IV. Numerical Result

We consider a simulation scenario where there are $L = 400$ single-antenna APs and $K = 100, 200, 300, 400$ UEs to demonstrate the performance of the proposed framework under different number of UEs from $L > K$ to $L \approx K$. The APs and UEs are uniformly distributed in a $2\times$ 2 km^2 wrapped-around area. The propagation model is the same as the one used in [8] but the APs are deployed 10 meters above the UEs. We assume the maximum traffic load imposed by an AP is $\mathcal{S}_l = 10$. Each UE transmits at full power with p_k=100 mW and 20MHz bandwidth. δ is set to as low as 10^{-4} to explore as many UEs in the exploration stage. TDD operation is considered with $\tau_c = 200$ blocks. The number of pilot τ_p is obtained from the graph-coloring based pilot assignment and the rest $\tau_u = \tau_c - \tau_p$ blocks are used for UL data transmission.

A. Uplink

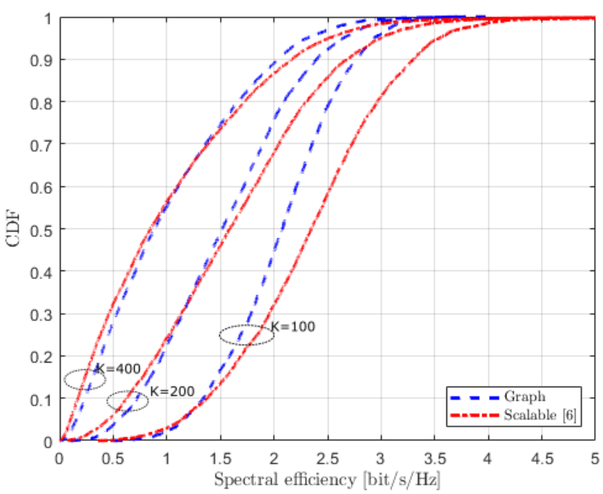

Fig. 4. Uplink Spectral Efficiency CDF

Fig. 4 shows the CDF of the uplink spectral efficiency of the proposed framework and the benchmark algorithm [6]. The benchmark has a fixed number of $\tau_p = 10$, whereas the proposed framework has a different value of τ_p depends on the number of UEs as shown in Fig. 6. The number of UEs served in the benchmark algorithm is limited by the number of pilots i.e., $|\mathcal{D}_l| \leqslant \tau_p = 10$ and the number of UEs served in the proposed framework is limited by the maximum traffic load imposed by an AP i.e., $|\mathcal{D}_l| \leqslant \mathcal{S}_l = 10$. The spectral efficiency is smaller as the number of UEs grow bigger because the average number of APs serving a UE decreases. Under the same number of served UEs limitation, the proposed framework outperforms the benchmark algorithm in terms of 95%-likely per-user SE when K is above 100.

Fig. 5 shows the 95%-likely per-user SE under different value of K. When the number of $K = 400$, the graph-based framework outperforms the benchmark algorithm 54% and when $K = 100$ the graph-based framework outperforms the benchmark algorithm 1.7%. This shows that the graph-based

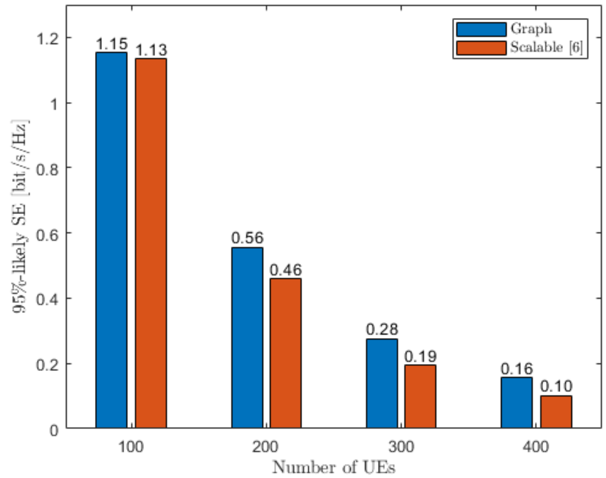

Fig. 5. 95%-likely per-user SE

framework provides a better 95%-likely per-user SE as the number of UEs grow bigger especially in the extreme case where $K \approx L$.

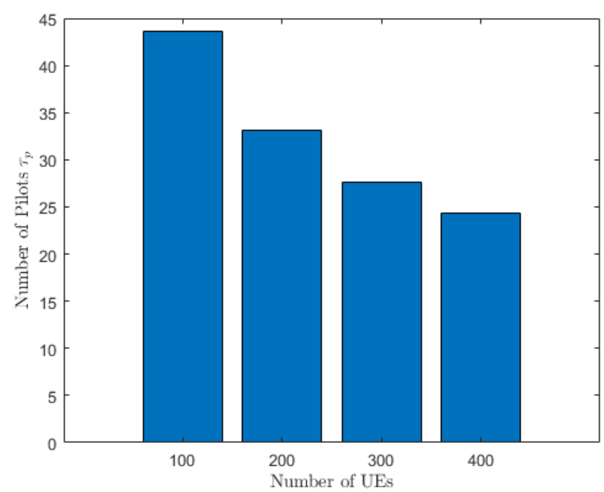

Fig. 6. Number of Pilots required for graph-based framework

Unlike the benchmark algorithm which has a fixed number of τ_p across a different number of UEs, graph-based framework has different τ_p under different number of UEs as the number of τ_p is determined from the output of Dsatur algorithm. Fig. 6 shows the relationship between the number of pilots used by graph-based framework versus the number of UEs. As the number of UEs grow bigger to L, the number of pilots becomes lower. This is because the average number of APs serving a UE is higher and that leads to more overlapping clusters. The higher number of overlapping clusters leads to more number of required colors as the adjacent vertices must not have the same color.

Equation (8) describes a trade-off between the number of pilots and the number of blocks used for uplink data transmission. When the number of pilots is higher, the SINR will be higher. Normally higher SINR leads to better SE, but note that the number of τ_c is limited, and we have the $\tau_c - \tau_p$ left for the usage of data transmission. If we set the number of τ_p too high, there is no room for data transmission and that high SINR becomes useless. When the number of UEs is a lot but not massive, e.g., 100 UEs, the high number of pilots makes the number of τ_u low, and the increase in SE due to high SINR can not keep up with the decrease in SE due to low τ_u, thus making the SE decreases. Hence, the proposed graph-based framework provides the best performance for the case when the number of UEs is massive, i.e., $K \approx L$.

B. Complexity Analysis

UE k has L_k AP candidates within its preference list and AP l has K_l UE candidates within its preference list. The worst complexity order of the anchoring stage is given by $\mathcal{O}(\sum_{k=1}^{K} L_k \log L_k + \sum_{l=1}^{L} K_l \log K_l)$. In the UE exploration stage, assuming the worst case scenario when all APs are not associated with any UE, AP l must explore K_l UEs which gives $\mathcal{O}(\sum_{l=1}^{L} K_l)$ complexity order. Moreover, the APs must sort the explored UEs with complexity of $\mathcal{O}(\sum_{l=1}^{L} K_l \log K_l)$. Finally, the complexity of graph coloring problem is on the order of $\mathcal{O}(K^2)$ which is better compared to exhaustive search scheme – the optimal scheme – which has complexity of $\mathcal{O}(\tau_p^K)$ resulting to a more scalable scheme when $K \rightarrow \infty$. Overall, the complexity of the proposed framework is on the order of $\mathcal{O}(K^2 + \sum_{k=1}^{K} L_k \log L_k + \sum_{l=1}^{L} K_l(\log K_l + 1))$.

V. Conclusion

In this paper, we have proposed a framework to form a cluster and assign pilots for UEs by considering the traffic limit imposed by each AP. The proposed framework outperforms the benchmark in terms of 95%-likely per-user SE when there are massive number of UEs. The complexity of the proposed framework is fairly low compared to the optimal solution. Hence this could be an interesting framework to use for user-centric cell-free massive MIMO scenario when $K \approx L$.

References

1 H. Q. Ngo, A. Ashikhmin, H. Yang, E. G. Larsson and T. L. Marzetta, ”Cell-Free Massive MIMO: Uniformly great service for everyone,” in IEEE 16th International Workshop on Signal Processing Advances in Wireless Communications (SPAWC), Stockholm, 2015.

2 E. Björnson and L. Sanguinetti, ”Making Cell-Free Massive MIMO Competitive With MMSE Processing and Centralized Implementation,” IEEE Transactions on Wireless Communications, vol. 19, no. 1, pp. 77-90, 2020.

3 H. Q. Ngo, A. Ashikhmin, H. Yang, E. G. Larsson and T. L. Marzetta, ”Cell-Free Massive MIMO Versus Small Cells,” IEEE Transactions on Wireless Communications, vol. 16, no. 3, pp. 1834 - 1850, 2017.

4 S. Buzzi and C. D'Andrea, ”Cell-free massive MIMO: User-centric approach,” IEEE Communications Letters, vol. 6, no. 6, p. 706–709, 2017.

5 H. Q. Ngo, L.-N. Tran, T. Q. Duong, M. Matthaiou and E. G. Larsson, ”On the Total Energy Efficiency of Cell-Free Massive MIMO,” IEEE Transactions on Green Communications and Networking, vol. 2, no. 1, pp. 25-39, 2018.

6 E. Björnson and L. Sanguinetti, ”Scalable Cell-Free Massive MIMO Systems,” IEEE Transactions on Communications, vol. 68, no. 7, pp. 4247 - 4261, 2020.

7 S. Chen, J. Zhang, E. Björnson, J. Zhang and B. Ai, ”Structured Massive Access for Scalable Cell-Free Massive MIMO Systems,” IEEE Journal on Selected Areas in Communications (Early Access), 2020.

8 E. Björnson, J. Hoydis and L. Sanguinetti, Massive MIMO Networks: Spectral, Energy, and Hardware Efficiency, nowpublishers, 2017.

9 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 Transactions on Vehicular Technology, vol. 67, no. 5, pp. 4440-4453, 2018.

10 D. Gale and L. Shapley, ”College admissions and the stability of marriage,” The American Mathematical Monthly, vol. 69, no. 1, pp. 9-15, 1962.

11 D. Brélaz, ”New methods to color the vertices of a graph,” Communications of the ACM, vol. 22, no. 4, p. 251–256, 1979.

12 H. Liu, J. Zhang, S. Jin and B. Ai, ”Graph Coloring Based Pilot Assignment for Cell-Free Massive MIMO Systems,” IEEE Transactions on Vehicular Technology, vol. 69, no. 8, pp. 9180 - 9184, 2020.