# Semi-Dynamic User-Specific Clustering for Downlink Cloud Radio Access Network

Dong Liu, Shengqian Han, Chenyang Yang, and Qian Zhang

***Abstract*—This paper studies user-specific clustering for downlink cloud radio access network (C-RAN), where a central unit, connected to all base stations (BSs) via limited-capacity backhaul links, coordinates the BSs to form cooperative clusters for every user. By taking into account the training overhead for channel estimation in C-RAN, we design the clustering scheme aimed at maximizing the average net throughput of the network subject to the constraint on backhaul capacity, where a hybrid coordinated multi-point (CoMP) transmission mode is considered. The proposed clustering scheme can be operated in a semi-dynamic manner merely based on large-scale channel information, has low computational complexity, and performs close to the optimal scheme found by exhaustive searching. Under two special cases where the backhaul capacity are very stringent and unlimited, the proposed scheme is then tailored for pure coordinated beamforming (CB) mode and pure joint transmission (JT) mode to further reduce the clustering complexity. Simulation results show that the proposed semi-dynamic clustering schemes are superior to the dynamic clustering scheme due to the reduction of required training overhead.**

***Index Terms*—Cloud radio access network (C-RAN), coordinated multi-point (CoMP), user-specific clustering, semi-dynamic, backhaul capacity.**

## I. INTRODUCTION

The telecom industry has been witnessing a traffic explosion in recent years, and has reached a broad consensus on the strong continuation of this trend for the next decade. To meet the ever demanding expectations of mobile broadband users in fifth-generation (5G), network densification and cloud radio access network (C-RAN) are recognized as two key enabling technologies [?]. By centralizing the baseband processing resources of all base stations (BSs) into a super resource pool at central unit (CU) and incorporating coordinated multi-point (CoMP) transmission techniques, C-RAN can significantly improve the system performance with affordable cost [?].

CoMP transmission can be generally divided into coordinated beamforming (CB) and joint transmission (JT). With CB, each BS only transmits the data intended for the user equipments (UEs) in its own cell, and forms beams to reduce interference to the UEs in neighboring cells. With JT, the cooperating BSs jointly transmit data to the UEs so that inter-cell interference (ICI) can be converted into useful signals. In general, JT outperforms CB when the CU can perfectly share the data of all UEs to the BSs [?]. However, if existing backhaul links are employed, which are capacity-limited [?, ?], JT may become inferior to CB [?]. A hybrid CoMP scheme switching between CB and JT was proposed in [?] for downlink CoMP transmission, which shows evident performance gain over the pure JT and pure CB modes.

In C-RAN systems, more channel information should be available to facilitate CoMP transmission compared with traditional single-cell transmission systems, i.e., Non-CoMP systems, which will cause larger training overhead [?]. The overhead increases with the number of cooperating BSs, and may even counteract the performance gain of CoMP transmission [?]. Considering that a UE choosing faraway BSs as its coordinated BSs will gain little in performance but with the penalty of increasing training overhead and complexity, CoMP is usually implemented within a cluster consisting of a limited number of BSs [?, ?, ?, ?]. The clustering schemes proposed in [?] and [?] form non-overlapped clusters based on geometrical information or instantaneous channel state information (CSI). These approaches are easy for implementation, but the UEs located at the cluster edge still suffer from severe interference from nearby clusters. The problem can be solved with user-specific clustering schemes as studied in [?] and [?], where each BS may belong to different clusters simultaneously, which inevitably leads to overlapped clusters for different UEs with different channel conditions.

However, acquiring instantaneous CSI for the existing user-specific clustering schemes will introduce large training overhead, which causes the performance degradation of CoMP systems as mentioned before. User-specific clustering based on large-scale channel information was studied in our preliminary work [?], where only pure CB mode was considered under the assumption of unlimited-capacity backhaul.

In this paper, we strive to study user-specific clustering schemes for downlink C-RAN to maximize the net system throughput by taking into account the training overhead for CSI acquisition. The main contributions of this paper are summarized as follows:

- We derive the closed-form expression of the asymptotical average data rate of UEs in high signal-to-noise ratio (SNR) regime, where a hybrid CoMP transmission is introduced considering that different kinds of backhaul links may be employed in C-RAN. Then, a weak interference estimation mechanism is provided to improve the accuracy of the asymptotical results for general SNRs.
- We design a low-complexity semi-dynamic user-specific clustering scheme merely based on large-scale channel

This work was supported by the National High Technology Research and Development Program of China (No. 2014AA01A703).



The authors are with the School of Electronics and Information Engineering, Beihang University, Beijing, China (e-mail: {dliu, sqhan, cyyang}@buaa.edu.cn; qianzhang@ee.buaa.edu.cn). S. Han is the corresponding author of this article.

information subject to the constraint on backhaul capacity. We proceed to study two special cases where the backhaul capacity is very stringent and is unlimited, respectively, under which the proposed scheme is tailored for pure CB and pure JT to further reduce the clustering complexity. Simulation results show that the proposed low-complexity semi-dynamic user-specific clustering schemes perform close to the optimal solution found by exhaustive searching, and outperform the optimal dynamic clustering scheme due to the reduction in the training overhead.

The remainder of the paper is organized as follows. In Section II, we present the system model. In Section III, the hybrid CoMP transmission under limited-capacity backhaul is introduced and the semi-dynamic user-specific clustering scheme is proposed. The proposed clustering scheme is then redesigned for pure CB and pure JT in Section IV. Simulation results are provided in Section V, and the conclusions are drawn in Section VI.

## II. System Model

Consider a downlink C-RAN consisting of $N_b$ BSs each equipped with $N_t$ antennas and connected with the CU via a backhaul link. The number of users scheduled by each BS is less than $N_t$, and $N_u$ users are scheduled in total in the network. Each scheduled UE is served by a cluster of BSs with a hybrid mode, where several master BSs send data to the UE and several coordinated BSs avoid interference to the UE.[1] The hybrid CoMP mode can be degenerated to three special cases as follows. When a UE has only one master BS and no less than one coordinated BS, the UE is served with pure CB. When a UE has multiple master BSs and no coordinated BS, the UE is served with pure JT. When a UE has one master BS and no coordinated BS, the UE is served with Non-CoMP, i.e., the UE receives the desired signal from one master BS and suffers interference from all the other BSs. An example of the considered C-RAN system is shown in Fig. **??**.

Fig. 1. Example of the considered C-RAN. A solid arrow stands for the connection between a UE and its master BS, and a dash arrow represents the connection between a UE and its coordinated BS. For instance, the master BSs of $\text{UE}_1$ include $\text{BS}_1$, $\text{BS}_5$ and $\text{BS}_6$, the coordinated BSs of $\text{UE}_7$ include $\text{BS}_1$ and $\text{BS}_6$, the served UEs of $\text{BS}_1$ include $\text{UE}_1$ and $\text{UE}_4$, and the coordinated UEs of $\text{BS}_3$ include $\text{UE}_2$ and $\text{UE}_4$.

Fig. 2. Illustration of the considered three-stage transmission strategy.

We consider time division duplex (TDD) systems with a three-stage transmission strategy, which is illustrated in Fig. **??**. In the first stage, each UE measures and reports the large-scale fading gains to the nearest BS (step $(a_1)$), then each BS sends the large-scale fading gains to the CU (step $(b_1)$), at which the clusters are formed for all UEs based on those reported large-scale fading gains (step $(c_1)$). In the second stage, each UE is informed to send uplink training sequences for CSI acquisition (step $(a_2)$), with which each BS estimates the downlink channels and computes the CoMP precoders for downlink transmission (step $(b_2)$). In the third stage, the CU shares the data of UEs to the BSs according to the clustering results obtained in the first stage (step $(a_3)$), and then the BSs transmit the downlink data (step $(b_3)$). Since the large scale channel information including user location and shadowing changes slowly, the interval of the first stage can be relatively large compared with the intervals of the second and third stages.

[1]In the sequel, the BSs that send signal to $\text{UE}_u$ are called the master BSs of $\text{UE}_u$, the BSs that avoid interference to $\text{UE}_u$ are called the coordinated BSs of $\text{UE}_u$, the UEs that receive data from $\text{BS}_b$ are called the served UEs by $\text{BS}_b$, and the UEs to which $\text{BS}_b$ avoids interference are called the coordinated UEs by $\text{BS}_b$. An example of the relationship is shown in Fig. **??**.

### A. Signal Model

To describe the clustering for different CoMP modes in a unified framework, we introduce *transmission matrix* $\mathbf{S} = [s_{ub}]_{N_u \times N_b}$ and *coordination matrix* $\mathbf{C} = [c_{ub}]_{N_u \times N_b}$ to reflect the relationship of a UE with its master BSs and coordinated BSs, respectively, whose elements are either 0 or 1. Specifically, if $\text{BS}_b$ is a master BS of $\text{UE}_u$, then $s_{ub} = 1$; otherwise, $s_{ub} = 0$. If $\text{BS}_b$ is a coordinated BS for $\text{UE}_u$, then $c_{ub} = 1$; otherwise, $c_{ub} = 0$. After the transmission matrix $\mathbf{S}$ and coordination matrix $\mathbf{C}$ are found, the clusters for all UEs are obtained.

Denote $\mathcal{T}_u = \{b | s_{ub} = 1\}$ and $\mathcal{P}_u = \{b | c_{ub} = 1\}$ as the set of the indices of the master BSs and the indices of the coordinated BSs of $\text{UE}_u$, respectively. Denote $\mathcal{S}_b = \{u | s_{ub} = 1\}$ and $\mathcal{I}_b = \{u | c_{ub} = 1\}$ as the sets of the indices of the UEs served by and coordinated by $\text{BS}_b$, respectively. Then, we can find that $\mathcal{C}_u = \mathcal{T}_u \cup \mathcal{P}_u$ is the set of the indices of all BSs in the cooperative cluster for $\text{UE}_u$, which is referred to as the cluster for $\text{UE}_u$.

Let $\alpha_{ub}$ and $\mathbf{h}_{ub} \in \mathbb{C}^{N_t \times 1}$ denote the large-scale channel gain and the small-scale channel vector from $\text{BS}_b$ to $\text{UE}_u$, respectively. The small-scale channel vectors are assumed as independent and identically distributed (i.i.d.) complex Gaussian random vectors with zero mean and covariance matrix $\mathbb{E}\{\mathbf{h}_{ub}\mathbf{h}_{ub}^H\} = \mathbf{I}$. Denoting the unit-norm precoding vector for $\text{UE}_u$ at $\text{BS}_b$ as $\mathbf{w}_{ub} \in \mathbb{C}^{N_t \times 1}$, then the received signal of $\text{UE}_u$ can be expressed as

$$y_u = \underbrace{\sum_{b \in \mathcal{T}_u} \alpha_{ub} \mathbf{h}_{ub}^H \mathbf{w}_{ub} \sqrt{p_{ub}} x_u}_{\text{desired signal}} + \sum_{m \neq u} \Bigg( \underbrace{\sum_{i \in \mathcal{T}_m \cap \mathcal{C}_u} \alpha_{ui} \mathbf{h}_{ui}^H \mathbf{w}_{mi} \sqrt{p_{mi}}}_{\text{intra-cluster interference}} + \underbrace{\sum_{j \in \mathcal{T}_m, j \notin \mathcal{C}_u} \alpha_{uj} \mathbf{h}_{uj}^H \mathbf{w}_{mj} \sqrt{p_{mj}}}_{\text{inter-cluster interference}} \Bigg) x_m + n_u, \quad (1)$$

where $p_{ub}$ is the transmit power allocated to $\text{UE}_u$ at $\text{BS}_b$, $x_u$ is the data symbol with unit variance destined to $\text{UE}_u$, and $n_u$ is the additive white Gaussian noise (AWGN) with zero mean and variance $\sigma_u^2$ at $\text{UE}_u$. In (**??**), the interference received by $\text{UE}_u$ includes two parts. The intra-cluster interference comes from the signal transmitted by $\text{BS}_i$ (i.e., a master BS of

UE$_m$ that is located inside the cluster of UE$_u$) to UE$_m$, which is received by UE$_u$ via channel $\mathbf{h}_{ui}$. The inter-cluster interference comes from the signal transmitted by BS$_j$ (i.e., a master BS of UE$_m$ that is located outside the cluster of UE$_u$) to UE$_m$, which is received by UE$_u$ via channel $\mathbf{h}_{uj}$.

The downlink signal-to-interference-plus-noise ratio (SINR) at UE$_u$ can be obtained from (**??**) as

$$\gamma_u = \frac{\left|\sum_{b\in\mathcal{T}_u}\sqrt{\lambda_{u,b}}\mathbf{h}_{ub}^H\mathbf{w}_{ub}\right|^2}{\sum_{m\neq u}\left|\sum_{i\in\mathcal{T}_m\cap\mathcal{C}_u}\sqrt{\lambda_{u,mi}}\mathbf{h}_{ui}^H\mathbf{w}_{mi}+\sum_{j\in\mathcal{T}_m,j\notin\mathcal{C}_u}\sqrt{\lambda_{u,mj}}\mathbf{h}_{uj}^H\mathbf{w}_{mj}\right|^2+\sigma_u^2} \triangleq \frac{S_u}{I_u+\sigma_u^2}, \tag{2}$$

where $S_u$ is the received power of desired signal, $I_u$ is the power of interference, $\lambda_{u,b} \triangleq \alpha_{ub}^2 p_{ub}$, $\lambda_{u,mi} \triangleq \alpha_{ui}^2 p_{mi}$, and $|\cdot|$ denotes the magnitude of a complex number. Then the downlink data rate of UE$_u$ can be expressed as

$$r_u = \log_2(1+\gamma_u). \tag{3}$$

In TDD systems, the channels can be estimated at the BS through uplink training by exploiting the reciprocity between uplink and downlink channels. With the minimum mean-square error (MMSE) criterion, the relationship between the estimated channel $\hat{\mathbf{h}}_{ub}$ and the true value of the channel $\mathbf{h}_{ub}$ satisfies [**?**]

$$\mathbf{h}_{ub} = \rho_{ub}\hat{\mathbf{h}}_{ub} + \mathbf{e}_{ub}, \tag{4}$$

where $\mathbf{e}_{ub} \sim \mathcal{CN}(0, \frac{1}{1+\eta_{ub}}\mathbf{I})$ is the channel estimation error, $\hat{\mathbf{h}}_{ub} \sim \mathcal{CN}(0,\mathbf{I})$, $\rho_{ub} = \frac{\sqrt{\eta_{ub}}}{\sqrt{1+\eta_{ub}}}$, $\eta_{ub} = \eta_{ub,0}\tau_{tr}$ is the equivalent average uplink receive SNR from UE$_u$ to BS$_b$. Herein, $\tau_{tr}$ is the number of uplink training symbols, and $\eta_{ub,0}$ is the average uplink receive SNR from UE$_u$ to BS$_b$.

### *B. Training Overhead*

For the considered three-stage transmission strategy, training may be required in the downlink precoding stage as well as the clustering stage if the clustering is based on instantaneous CSI. After taking into account the uplink training overhead, the net downlink data rate of UE$_u$ can be expressed as

$$R_u(\mathbf{S},\mathbf{C}) = (1-vT)r_u, \tag{5}$$

where $v = \frac{\tau_{tr}}{\tau}$ represents the percentage of resources taken by the uplink training of each UE in TDD systems, $\tau$ is the total number of symbols of each frame, and $T$ reflects the occupied uplink resources to ensure the orthogonality among the training signals of multiple UEs [**?**], where we assume that the training duration is smaller than the channel coherence time because CoMP is typically adopted in low-mobility environments. The impact of training overhead $v$ on the net downlink data rate is twofold. On one hand, increasing $v$ leads to accurate channel estimation (noting that $\eta_{ub} = \eta_{ub,0}\tau v$) and hence improves the data rate. On the other hand, however, a large value of $v$ wastes more system resources that decreases the net downlink data rate.

The value of $T$ depends on the formed clusters $\mathcal{C}_u$, $u = 1,\cdots,N_u$. When all the clusters are not overlapped,[2] only the training signals of the UEs located in the same cluster should be orthogonal, i.e., UE$_u$ and UE$_m$ should use orthogonal training signals if $\mathcal{C}_u = \mathcal{C}_m$. In this case, $T$ simply equals to the number of scheduled UEs in each cluster, which is $T = n\big(\bigcup_{b\in\mathcal{C}_u}(\mathcal{S}_b\cup\mathcal{I}_b)\big)$, where $n(\cdot)$ denotes the size of a set. When the clusters of different UEs are overlapped, the training signals of the UEs in different clusters served or coordinated by the same BS should also be orthogonal so that the BS can distinguish the UEs, i.e., UE$_u$ and UE$_m$ should use orthogonal training signals if $\mathcal{C}_u\cap\mathcal{C}_m \neq \varnothing$ (empty set). In overlapped clusters the value of $T$ can be computed by using the algorithm proposed in [**?**], which is briefly summarized in Appendix **??** by taking the clusters shown in Fig. **??** as an example for the reader's convenience.

If the clusters are formed statically based on geographical information or semi-dynamically based on large-scale channel information, then uplink training is only needed in the second stage for downlink precoding and the overhead is determined by the already formed clusters. By contrast, if the clusters are formed based on small scale channels, uplink training is required in the first stage before clustering, which generally leads to more training overhead for a large pool of master BSs and coordinated BSs, from which the clusters are selected.

## III. Semi-dynamic User-specific Clustering with Hybrid Mode

As previously mentioned, dynamically forming cooperative clusters based on small-scale fading channels yields frequently changing clusters and leads to large signaling overhead among BSs and UEs, making it infeasible in practical systems. Therefore, in the sequel, we propose to form cooperative cluster for each UE based on large-scale channels, aiming at maximizing the average rather than the instantaneous net downlink throughput. As a result, the proposed scheme can be implemented in a semi-dynamic manner.

The average net throughput of C-RAN system depends on the employed CoMP mode, which relies on the available backhaul capacity. It has been shown that pure JT or pure CB only works well when the backhaul capacity limitation is very loose [**?**] or very stringent [**?**], but both perform poorly in other cases. In this section, we first introduce a so-called hybrid CoMP mode, with which the average net throughput of C-RAN system is derived. Then, a semi-dynamic user-specific clustering scheme is proposed under limited-capacity backhaul.

### *A. Hybrid CoMP Mode*

In hybrid CoMP mode, each BS has multiple served UEs and multiple coordinated UEs, and the clusters for different UEs are generally overlapped, such that existing multiuser

[2] In non-overlapped clusters each BS only belongs to a single cluster, or in other words, if BS$_b$ is selected by UE$_u$ and UE$_m$ as a master or coordinated BS, then $\mathcal{C}_u = \mathcal{C}_m$ holds, i.e., UE$_u$ and UE$_m$ select the same cluster. By contrast, when the clusters of different UEs are overlapped, some BSs will be shared by multiple users, which belong to different clusters simultaneously.

multi-input multi-output (MU-MIMO) based CoMP precoders that regard a non-overlapped cluster as a super cell can not be applied, unless one considers the whole C-RAN as a cluster, which however is computationally prohibitive in practice. To decoupled the precoders of the BSs under overlapped clusters, the precoder of a BS is designed to avoid ICI to its coordinated UEs, and also to ensure that the signal sent to a served UE is co-phased with the signals from other master BSs of the UE for constructive combination. Under the principle of zero-forcing (ZF), which is with low complexity and widely used for CoMP transmission, the precoder in hybrid CoMP mode can be designed as follows.

Let $\text{BS}_b$ denote a master BS of $\text{UE}_u$. To avoid generating the interference to all coordinated UEs in $\mathcal{I}_b$ and all served UEs in $\mathcal{S}_b$ except $\text{UE}_u$, the ZF precoding vector for $\text{UE}_u$ at $\text{BS}_b$ is designed as

$$\mathbf{w}_{ub} = \frac{\Pi_{\bar{u}b}\hat{\mathbf{h}}_{ub}}{\left\|\Pi_{\bar{u}b}\hat{\mathbf{h}}_{ub}\right\|},\ b \in \mathcal{T}_u, \tag{6}$$

where $\Pi_{\bar{u}b} = \mathbf{I} - \hat{\mathbf{H}}_{\mathcal{K}_u,b}^H \left(\hat{\mathbf{H}}_{\mathcal{K}_u,b}\hat{\mathbf{H}}_{\mathcal{K}_u,b}^H\right)^{-1} \hat{\mathbf{H}}_{\mathcal{K}_u,b}$ is the null space of $\hat{\mathbf{H}}_{\mathcal{K}_u,b} \in \mathbb{C}^{n(\mathcal{K}_u)\times N_t}$, which is the estimated channel matrix from $\text{BS}_b$ to the UEs inside $\mathcal{K}_u$ with $\mathcal{K}_u = \mathcal{S}_b \cup \mathcal{I}_b - \{u\}$ denoting all served and coordinated UEs by $\text{BS}_b$ except $\text{UE}_u$. With the precoding vector in (**??**), the signals of $\text{UE}_u$ sent from all master BSs in $\mathcal{T}_u$ will be constructively combined if the CSI is perfect, because the equivalent channels $\mathbf{h}_{ub}^H\mathbf{w}_{ub}$ are all real-positive for $b \in \mathcal{T}_u$. Imperfect channel estimation will degrade the strength of the desired signal since the equivalent channels from multiple BSs are no longer perfectly co-phased.

Note that the hybrid CoMP mode degenerates to the pure CB or JT mode when $n(\mathcal{T}_u) = 1$ or $\mathcal{P}_u = \varnothing$. With the above ZF precoder, the overall number of UEs that are served and coordinated simultaneously by $\text{BS}_b$ is restricted by the number of antennas at $\text{BS}_b$, i.e., $M_b \triangleq n(\mathcal{S}_b) + n(\mathcal{I}_b) = \sum_{u=1}^{N_u}(s_{ub} + c_{ub}) \le N_t, b = 1, \cdots, N_b$.

Intuitively, as analyzed in [**?**], a BS should allocate more power to the users that have strong links to the BS in order to achieve higher sum rate. Further considering that we aim to propose semi-dynamic clustering schemes based on large-scale channel gains, the semi-dynamic power allocation given in [**?**] is employed, where $p_{ub}$ is set proportional to $\alpha_{ub}^2$, which is $p_{ub} = \frac{\alpha_{ub}^2}{\sum_{k\in\mathcal{S}_b}\alpha_{kb}^2}P$ for $u \in \mathcal{S}_b$ with $P$ denoting the transmit power of each BS. With $\{p_{ub}\}$ and $\{\mathbf{w}_{ub}\}$, we can compute the SINR of $\text{UE}_u$ based on (**??**) and then analyze the average net throughput of the C-RAN system.

### *B. Average Data Rate of Each User*

In this subsection we derive the average data rate of each user under hybrid CoMP mode, which will be used in the subsequent design of clustering methods. We first obtain the distributions of the signal and interference terms in the SINR given in (**??**). Then, we derive the asymptotical average data rate of $\text{UE}_u$ and obtain a closed-form expression in high SNR regime.

*1) Signal Term:*

**Proposition 1:** The signal term can be approximated as

$$S_u \approx \Big(\sum_{b\in\mathcal{T}_u}\sqrt{\lambda_{u,b}}\rho_{ub}\hat{\mathbf{h}}_{ub}^H\mathbf{w}_{ub}\Big)^2 \triangleq \Big(\sum_{b\in\mathcal{T}_u} S_{u,b}\Big)^2, \tag{7}$$

where $S_{u,b}$ follows Nakagami-$m$ distribution with parameters $\Omega_{ub} = K_b\lambda_{u,b}\rho_{ub}^2$ and $m_b = K_b\ \mathbb{G}(K_b, 1)$. Herein, $\lambda_{u,b} = \alpha_{ub}^2\frac{\alpha_{ub}^2 P}{\sum_{k\in\mathcal{S}_b}\alpha_{kb}^2}$ and $K_b = N_t - M_b + 1$.

*Proof:* See Appendix **??**. ■

From Proposition 1, we can see that the signal term $S_u$ is the square of the sum of Nakagami-$m$ random variables (RVs), $S_{u,b}$. However, the probability density function (PDF) of the sum of independent non-identically distributed Nakagami-$m$ RVs has no closed-form expression in general [**?**], which makes it difficult to derive the distribution of the signal term $S_u$. Fortunately, the distribution of the sum of independent non-identically distributed Nakagami-$m$ RVs can be well approximated by a Nakagami-$m$ distribution over a wide range of parameters via matching the first two moments (i.e., the moment matching method) [**?**, **?**]. Since the square of a Nakagami-$m$ distributed RV follows Gamma distribution. Therefore, $S_u$ can be well approximated by a Gamma distributed RV denoted as $\hat{S}_u \sim \mathbb{G}(\hat{k}_u, \hat{\theta}_u)$ via matching the first two moments of $S_u$, which will be verified through simulations later.

**Proposition 2:** By matching the first two moments of $S_u$ and $\hat{S}_u$, the parameters $\hat{k}_u$ and $\hat{\theta}_u$ can be derived as

$$\hat{\theta}_u = \frac{\mathbb{E}\{S_u^2\} - \mathbb{E}^2\{S_u\}}{\mathbb{E}\{S_u\}}, \tag{8a}$$

$$\hat{k}_u = \frac{\mathbb{E}^2\{S_u\}}{\mathbb{E}\{S_u^2\} - \mathbb{E}^2\{S_u\}}, \tag{8b}$$

where $\mathbb{E}\{S_u\}$ and $\mathbb{E}\{S_u^2\}$ are given in Appendix **??**.

*Proof:* See Appendix **??**. ■

The value of $\hat{k}_u$ may not be integers, which makes it difficult to obtain a closed-form expression of the average data rate. To tackle this problem, we further approximate $\hat{S}_u$ by rounding off $\hat{k}_u$. Let $\tilde{S}_u \sim \mathbb{G}(\tilde{k}_u, \tilde{\theta}_u)$ denote the resultant Gamma distributed RV, where $\tilde{k}_u$ is the nearest integer to $\hat{k}_u$, and $\tilde{\theta}_u = \frac{\hat{k}_u\hat{\theta}_u}{\tilde{k}_u}$ guaranties the match of the first moment between $\hat{S}_u$ and $\tilde{S}_u$.

*2) Interference term:* The interference term $I_u$ is

$$\sum_{m\neq u}\Big|\sum_{i\in\mathcal{T}_m\cap\mathcal{C}_u}\sqrt{\lambda_{u,mi}}\mathbf{h}_{ui}^H\mathbf{w}_{mi} + \sum_{j\in\mathcal{T}_m, j\notin\mathcal{C}_u}\sqrt{\lambda_{u,mj}}\mathbf{h}_{uj}^H\mathbf{w}_{mj}\Big|^2$$
$$\triangleq \sum_{m\neq u}\Big|J_{um}^{\text{Intra}} + J_{um}^{\text{Inter}}\Big|^2, \tag{9}$$

where $J_{um}^{\text{Intra}}$ and $J_{um}^{\text{Inter}}$ are the intra-cluster interference and inter-cluster interference experienced at $\text{UE}_u$ from the signals sent to $\text{UE}_m$, respectively.

The first term $J_{um}^{\text{Intra}}$ can be derived as

$$J_{um}^{\text{Intra}} = \sum_{i\in\mathcal{T}_m\cap\mathcal{C}_u}\sqrt{\lambda_{u,mi}}(\rho_{ui}\hat{\mathbf{h}}_{ui} + \mathbf{e}_{ui})^H\mathbf{w}_{mi}$$
$$= \sum_{i\in\mathcal{T}_m\cap\mathcal{C}_u}\sqrt{\lambda_{u,mi}}\mathbf{e}_{ui}^H\mathbf{w}_{mi}, \tag{10}$$

where $\hat{\mathbf{h}}_{ui}^H\mathbf{w}_{mi} = 0$ due to ZF precoding, and $\mathbf{e}_{ui}^H\mathbf{w}_{mi} \sim \mathcal{CN}(0, \frac{1}{1+\eta_{ui}})$ for $i \in \mathcal{T}_m \cap \mathcal{C}_u$ since $\mathbf{e}_{ui}$ is a complex Gaussian vector and independent from the unit-norm vector $\mathbf{w}_{mi}$. Thus, $J_{um}^{\text{Intra}} \sim \mathcal{CN}(0, \sum_i \frac{\lambda_{u,mi}}{1+\eta_{ui}})$.

The second term is

$$J_{um}^{\text{Inter}} = \sum_{j\in\mathcal{T}_m, j\notin\mathcal{C}_u} \sqrt{\lambda_{u,mj}}\mathbf{h}_{uj}^H\mathbf{w}_{mj}. \tag{11}$$

Due to the mutual independence between $\mathbf{h}_{uj}^H$ and $\mathbf{w}_{mj}$, $\sqrt{\lambda_{u,mj}}\mathbf{h}_{uj}^H\mathbf{w}_{mj} \sim \mathcal{CN}(0, \lambda_{u,mj})$, we have $J_{um}^{\text{Inter}} \sim \mathcal{CN}(0, \sum_j \lambda_{u,mj})$. Therefore, $|J_{um}^{\text{Inter}} + J_{um}^{\text{Intra}}|^2$ follows exponential distribution with mean $\lambda_{um} \triangleq \sum_i \frac{\lambda_{u,mi}}{1+\eta_{ui}} + \sum_j \lambda_{u,mj}$, and the interference term $I_u$ is the sum of exponential distributed RVs, whose PDF can be obtained from [?] as

$$f_{I_u}(x) = \sum_{m\neq u} \delta_{um} e^{-\frac{x}{\lambda_{um}}}, \tag{12}$$

where $\delta_{um} = \frac{1}{\lambda_{um}} \prod_{m'\neq m} \frac{\lambda_{um}}{\lambda_{um}-\lambda_{um'}}$ and $\lambda_{um'} \neq \lambda_{um}$ for $m' \neq m$.

Even with the distributions of $S_u$ and $I_u$, the noise term $\sigma_u^2$ still makes it hard to obtain a closed-form expression of the average data rate. In the following proposition, an asymptotic result in high SNR regime (i.e., $\sigma_u^2 \to 0$) is derived, which can be used as an approximated average date rate of $\text{UE}_u$ and is accurate for high SNRs.

**Proposition 3:** When $\sigma_u^2$ is ignorable, the average data rate of $\text{UE}_u$ can be approximated as

$$\bar{r}_u \approx \sum_{m\neq u} \xi_{um}\Big( -\ln\tilde{\zeta}_{um} + \sum_{k=1}^{\tilde{k}_u-1}\frac{1}{k} + \frac{(-\tilde{\zeta}_{um})^{\tilde{k}_u}\ln\tilde{\zeta}_{um}}{(1-\tilde{\zeta}_{um})^{\tilde{k}_u}} - \sum_{k=1}^{\tilde{k}_u-1}\binom{\tilde{k}_u-1}{k}\frac{(1-\tilde{\zeta}_{um}^k)(-\tilde{\zeta}_{um})^{\tilde{k}_u-k}}{k(1-\tilde{\zeta}_{um})^{\tilde{k}_u}}\Big), \tag{13}$$

where $\xi_{um} = \frac{1}{\ln 2}\prod_{m'\neq m}\frac{\lambda_{um}}{\lambda_{um}-\lambda_{um'}}$, $\tilde{\zeta}_{um} = \frac{\lambda_{um}}{\tilde{\theta}_u}$, and $\tilde{k}_u$ and $\tilde{\theta}_u$ are defined after Proposition 2.

*Proof:* See Appendix **??**. ■

Note that the parameters $\xi_{un}$, $\tilde{\zeta}_{um}$, $\tilde{k}_u$ and $\tilde{\theta}_u$ in (**??**) only depend on the large-scale channel gains $\alpha_{ub}$, $u = 1, \cdots, N_u$ and $b = 1, \cdots, N_b$. Moreover, the expression in (**??**) only consists of arithmetic operations of real numbers, which can be computed with low complexity.

### *C. Average Net Throughput Under Limited-Capacity Backhaul*

Given the transmission matrix $\mathbf{S}$ and the coordination matrix $\mathbf{C}$ as well as the semi-dynamic power allocation specified in Section **??**, the average downlink data rates of all UEs under limited-capacity backhaul, denoted by $\bar{\mathbf{r}}^* = (\bar{r}_1^*, \cdots, \bar{r}_{N_u}^*)$, can be obtained by solving the following linear programming (LP) problem,

$$\max_{\bar{\mathbf{r}}^*} \quad (\bar{\mathbf{r}}^*)^T\mathbf{1} \tag{14a}$$

$$s.t. \quad \mathbf{S}^T\bar{\mathbf{r}}^* \leq C\cdot\mathbf{1} \tag{14b}$$

$$\bar{\mathbf{r}}^* \leq \bar{\mathbf{r}}, \tag{14c}$$

where $\bar{\mathbf{r}} = (\bar{r}_1, \cdots, \bar{r}_{N_u})$ consists of the average downlink data rate of each UE without considering the backhaul capacity constraint, $C$ is the normalized capacity of each backhaul link by the transmission bandwidth, and $\mathbf{1} \in \mathbb{R}^{N_b\times 1}$ is an all-one vector. Constraint (**??**) indicates that the average traffic load of each backhaul link is limited by its capacity, where the entry of the vector $\mathbf{S}^T\bar{\mathbf{r}}^*$ denotes the sum data rate transmitted by each BS,[3] and constraint (**??**) indicates that the average downlink data rate of each UE under limited-capacity backhaul is no larger than that without backhaul capacity constraint. Problem (**??**) can be solved efficiently by standard convex optimization algorithms [**?**].

Then, the average net throughput of the C-RAN system under limited-capacity backhaul can be expressed as

$$\bar{R}^*(\mathbf{S}, \mathbf{C}) = \sum_{u=1}^{N_u}(1-vT)\bar{r}_u^*, \tag{15}$$

where $\bar{r}_u^*$ for $u = 1, \cdots, N_u$ are the optimal solution to problem (**??**).

### *D. Measurement Set and Weak Interference Estimation Mechanism*

It can be found from (**??**), (**??**) and (**??**) that the obtained average net throughput only depends on the large-scale channel gains $\alpha_{ub}$, $u = 1, \cdots, N_u$ and $b = 1, \cdots, N_b$, which can be measured at each UE by averaging the channels in multiple time slots and then be reported to the UE's nearest BS. Then, the CU can gather the large-scale fading gains of all UEs from all BSs, based on which the cooperative clusters can be formed.

In fact, reporting all large-scale fading gains is unnecessary because it benefits little for a UE to choose faraway BSs as its master or coordinated BSs. More importantly, it can be seen later that the complexity of the clustering schemes proposed in the sequel will increase quadratically with the number of reported large-scale fading gains. Therefore, we restrict that $\text{UE}_u$ only reports the large-scale channel gains from the BSs in a so-called "measurement set", denoted by $\mathcal{F}_u$. Intuitively, the measurement set of each user should contain the BSs with strong channel gains, which can be determined with the following two simple ways.

1) Fixed-size measurement set: Let $N_f \triangleq n(\mathcal{F}_u)$ denote the size of $\mathcal{F}_u$, $u = 1, \ldots, N_u$. Then, each UE only reports $N_f$ strongest large-scale channels.
2) Threshold-based measurement set: Each UE selects the BSs with relatively large channel gains, e.g., with the method in [**?**],

$$\mathcal{F}_u = \left\{ b \,\middle|\, \frac{\alpha_{ub}^2}{\max\left\{\alpha_{uj}^2\right\}_{j=1}^{N_b}} > \beta \right\},$$

where $0 < \beta \leq 1$ is a predefined threshold. The resultant measure sets may have different sizes for different UEs.

[3]Recalling that the ZF precoder is computed at each BS based on its own channels, it is not necessary to share CSI via backhaul links. Therefore, only the backhauling traffic caused by data sharing is taken into account.

The problem of introducing measurement set is that the CU only has a part of large-scale fading gains so that we cannot compute the average data rate based on Proposition 3 directly. A simple way to cope with this problem is to set the unknown large-scale fading gains as zeros. However, although the interference from each of the BSs outside the measurement set is very weak when the measurement set is large, the total interference from those BSs may not be ignorable. If the CU simply ignores the interference when selecting cluster for each UE, it will overestimate the net throughput of the network, which may lead to improper clustering results and degrade the system performance.

Next, we propose a weak interference estimation mechanism to address the problem. First, except for reporting the large-scale fading gains from the BSs in measurement set, we also let each UE measure and report its total average received power, which is a scalar feedback and can be easily implemented via the reference signal receiving power (RSRP) measurement and the reporting mechanism in long term evolution (LTE) systems. Denote the total average received power of $\mathrm{UE}_u$ as $Q_u$. Then, the total average interference power outside the measurement set plus the noise power for $\mathrm{UE}_u$ can be obtained as

$$\sum_{i\notin\mathcal{F}_u} P\alpha_{ui}^2+\sigma_u^2=Q_u-\sum_{b\in\mathcal{F}_u} P\alpha_{ub}^2. \tag{16}$$

Assume that the distances between $\mathrm{UE}_u$ and the BSs generating weak interference follow uniform distribution. Then, with the total average power of weak interference and noise, we can estimate the large-scale fading gains $\alpha_{ui}^2$ for the weak interference channels with $i\notin\mathcal{F}_u$ as

$$\alpha_{ui}^2=\frac{\theta_{ui}^{-\nu}}{\sum\limits_{j\notin\mathcal{F}_u}\theta_{uj}^{-\nu}}\cdot\frac{1}{P}\Big(Q_u-\sum_{b\in\mathcal{F}_u}P\alpha_{ub}^2\Big), \tag{17}$$

where $\nu$ denotes the path loss exponent, and $\{\theta_{uj}\}$ are i.i.d. random variables following the uniform distribution $[0,1]$, which denote the normalized distances from $\mathrm{BS}_j$ to $\mathrm{UE}_u$ within $[0,1]$ because it is shown in (**??**) that any scaling over $\{\theta_{uj}\}$ will not affect the value of $\alpha_{ui}^2$. In Section VI, we verify that with such an estimation, the CU can compute the average net throughput more accurately than directly using (**??**). This is because the weak interference estimation mechanism takes into account the noise power in $Q_u$, while the noise is set as zero in the asymptotical result given in (**??**).

The measurement and reporting of the large-scale channel information from the UEs cause negligible extra overhead on training or signaling, since these information changes slowly and thus the measurement and the reporting interval can be very large.

### *E. Semi-dynamic User-specific Clustering in Hybrid CoMP Mode*

After obtaining the large-scale channel gains, the semi-dynamic user-specific clustering can be optimized aimed at maximizing the average net throughput of the network under limited-capacity backhaul. The clustering in the hybrid CoMP mode is to jointly design the transmission matrix $\mathbf{S}$ and coordination matrix $\mathbf{C}$, which can be formulated as,

$$\max_{\mathbf{S},\mathbf{C}} \quad \bar{R}^*(\mathbf{S},\mathbf{C}) \tag{18a}$$
$$s.t. \quad c_{ub}s_{ub}=0 \tag{18b}$$
$$s_{ub}=0,\quad b\notin\mathcal{F}_u \tag{18c}$$
$$c_{ub}=0,\quad b\notin\mathcal{F}_u \tag{18d}$$
$$\sum_{u=1}^{N_u}(c_{ub}+s_{ub})\le N_t, \tag{18e}$$

where (**??**) indicates that a BS cannot be a master BS and a coordinated BS for a UE at the same time, (**??**) and (**??**) indicates that the master BSs and the coordinated BSs for $\mathrm{UE}_u$ should be selected from its measurement set, and (**??**) indicates that the total number of served and coordinated UEs by $\mathrm{BS}_b$ should be less than the number of antennas at $\mathrm{BS}_b$.

Since the possible values of $(s_{ub},c_{ub})$ can be $(0,1),(1,0)$, or $(0,0)$ considering the constraint in (**??**), the complexity of finding the optimal transmission matrix $\mathbf{S}$ and coordination matrix $\mathbf{C}$ by exhaustive searching is $\mathcal{O}(3^{N_uN_f})$, which is of prohibitive complexity. In the following, we propose a sub-optimal low-complexity algorithm to design $\mathbf{S}$ and $\mathbf{C}$, where the transmission links and cooperative links are successively selected from the initialization of Non-CoMP until the average net throughput of the network under limited-capacity backhaul stops increasing. The algorithm can be briefly summarized as follows.

1) **Initialization:** Initialize $\mathbf{S}^{(0)}$ by letting each UE select the BS with the largest average channel gain as a master BS, and let $\mathbf{C}^{(0)}=\mathbf{0}$. Compute $\bar{R}_{\max}^{*(0)}=\bar{R}^*(\mathbf{S}^{(0)},\mathbf{C}^{(0)})$, which is the average net throughput under limited-capacity backhaul when all UEs are served in Non-CoMP mode. Set $i=1$.
2) **Iteration:**
   a) Find the optimal transmission link or cooperative link maximizing the average net throughput by solving the following problem over $2(N_uN_f-i)$ possible candidates,

$$\max_{\mathbf{D}^{(i)}} \quad \max\big\{\bar{R}^*(\mathbf{S}^{(i-1)}+\mathbf{D}^{(i)},\mathbf{C}^{(i-1)}),\; \bar{R}^*(\mathbf{S}^{(i-1)},\mathbf{C}^{(i-1)}+\mathbf{D}^{(i)})\big\} \tag{19a}$$
$$s.t. \quad \sum_{u=1}^{N_u}\sum_{b=1}^{N_b} d_{ub}^{(i)}\le 1,\; d_{ub}^{(i)}\in\{0,1\} \tag{19b}$$
$$d_{ub}^{(i)}=0,\quad b\notin\mathcal{F}_u \tag{19c}$$
$$d_{ub}^{(i)}c_{ub}^{(i-1)}=0 \tag{19d}$$
$$d_{ub}^{(i)}s_{ub}^{(i-1)}=0 \tag{19e}$$
$$\sum_{u=1}^{N_u}(c_{ub}^{(i-1)}+s_{ub}+d_{ub}^{(i)})\le N_t, \tag{19f}$$

   where $\mathbf{D}^{(i)}=[d_{ub}]_{N_u\times N_b}$ is a zero matrix except for one element which equals to "1" to denote the newly added link, constraint (**??**) indicates that the number of "1"s in $\mathbf{D}^{(i)}$ is no larger than 1, constraints (**??**), (**??**) and (**??**) indicate that each UE can only choose the newly added master BS or coordinated BS from its measurement set excluding the already selected BSs, and constraint (**??**) indicates that the total number of served and coordinated UEs at each BS should be limited by the number of its antennas. The searching space to solve this problem is $2(N_uN_f - i)$, which shrinks with the increase of iteration times $i$.

b) Denote $\mathbf{D}_{\text{opt}}^{(i)}$ as the optimal value of $\mathbf{D}^{(i)}$, and define $\bar{R}_{\max}^{*(i)} \triangleq \max\{\bar{R}^*(\mathbf{S}^{(i-1)}, \mathbf{C}^{(i-1)} + \mathbf{D}_{\text{opt}}^{(i)}), \bar{R}^*(\mathbf{S}^{(i-1)} + \mathbf{D}_{\text{opt}}^{(i)}, \mathbf{C}^{(i-1)})\}$.

c) If $\bar{R}^*(\mathbf{S}^{(i-1)}, \mathbf{C}^{(i-1)} + \mathbf{D}_{\text{opt}}^{(i)}) \geq \bar{R}^*(\mathbf{S}^{(i-1)} + \mathbf{D}_{\text{opt}}^{(i)}, \mathbf{C}^{(i-1)})$, then add a new cooperative link into the system, i.e., update $\mathbf{C}^{(i)} = \mathbf{C}^{(i-1)} + \mathbf{D}_{\text{opt}}^{(i)}$; otherwise, add a new transmission link into the system, i.e. update $\mathbf{S}^{(i)} = \mathbf{S}^{(i-1)} + \mathbf{D}_{\text{opt}}^{(i)}$.

d) With the increase of the number of cooperative links, the training overhead grows and the available antenna resources diminish, which finally leads to the reduction of the average net throughput. The iteration stops when $\bar{R}_{\max}^{*(i)} \leq \bar{R}_{\max}^{*(i-1)}$, and $\mathbf{S}^{(i-1)}$ and $\mathbf{C}^{(i-1)}$ are the final results of the transmission matrix and the coordination matrix, respectively.

The overall computational complexity of this algorithm is $\mathcal{O}(\sum_{i=0}^{N_uN_f} 2(N_uN_f - i)) = \mathcal{O}(N_u^2N_f^2)$, which is much less than exhaustive searching with $\mathcal{O}(3^{N_uN_f})$.

## IV. Semi-dynamic User-specific Clustering with Pure CB and Pure JT Modes

As described in Section II, the hybrid CoMP mode is a combination of two CoMP transmission, pure CB and pure JT. It is not hard to understand that if the backhaul capacity is very stringent, pure CB is the optimal CoMP transmission scheme since it has the least backhaul capacity requirement, while if the backhaul capacity is unlimited, pure JT is optimal since it can make full use of the data sharing capability of backhaul links. Compared to the hybrid CoMP mode, the clustering complexity can be further reduced in pure CB or pure JT because only coordinated BSs or master BSs need to be selected for each mode. In the sequel, the clustering methods for pure CB and pure JT are studied, respectively.

### *A. Semi-dynamic User-specific Clustering with Pure CB Mode*

The degeneration from hybrid CoMP to pure CB simplifies the signal model, which gives us the opportunity to improve the accuracy of the obtained average net throughput and also to further reduce the complexity of the proposed semi-dynamic user-specific clustering scheme.

*1) Average Net Throughput in Pure CB Mode:*

*a) Signal Term:* With pure CB, each UE has only one master BS. Then, we have $\mathcal{T}_u = \{b_u\}$ with $b_u$ denoting the master BS of UE$_u$ and the signal term in (**??**) reduces to $S_u \approx \lambda_{u,b_u}\rho_{ub_u}^2|\hat{\mathbf{h}}_{ub_u}^H\mathbf{w}_{ub}|^2$, which follows Gamma distribution $\mathbb{G}(k_u, \lambda_{u,b_u}\rho_{ub_u}^2)$ with integer $k_u = N_t - M_{b_u} + 1$. Since $S_u$ has only one term following exact Gamma distribution with integer parameter $k_u$, the Gamma approximation in Proposition 2 and the round-off approximation in Section **??** can be removed.

*b) Interference Term:* Since $n(\mathcal{T}_u) = 1$, the interference term in (**??**) can be simplified into

$$\begin{aligned} I_u &= \sum_{m\neq u, b_m\in\mathcal{C}_u} \left|\sqrt{\lambda_{u,mb_m}}\mathbf{h}_{ub_m}^H\mathbf{w}_{mb_m}\right|^2 + \sum_{k\neq u, b_k\notin\mathcal{C}_u} \left|\sqrt{\lambda_{u,kb_k}}\mathbf{h}_{ub_k}^H\mathbf{w}_{kb_k}\right|^2 \\ &\triangleq \sum_{m\neq u, b_m\in\mathcal{C}_u} I_{um}^{\text{Intra}} + \sum_{k\neq u, b_k\notin\mathcal{C}_u} I_{uk}^{\text{Inter}}, \end{aligned} \tag{20}$$

where $I_{um}^{\text{Intra}}$ and $I_{uk}^{\text{Inter}}$ are the intra-cluster interference power and inter-cluster interference power suffered by UE$_u$ from UE$_m$ and UE$_k$, respectively. The first term $I_{um}^{\text{Intra}}$ can be derived as,

$$\begin{aligned} I_{um}^{\text{Intra}} &= \left|\sqrt{\lambda_{u,mb_m}}(\rho_{ub_m}\hat{\mathbf{h}}_{ub_m} + \mathbf{e}_{ub_m})^H\mathbf{w}_{mb_m}\right|^2 \\ &= \lambda_{u,mb_m}\left|\mathbf{e}_{ub_m}^H\mathbf{w}_{mb_m}\right|^2, \end{aligned} \tag{21}$$

where $\hat{\mathbf{h}}_{ub_m}^H\mathbf{w}_{mb_m} = 0$ due to ZF precoding, and $\mathbf{e}_{ub_m}^H\mathbf{w}_{mb_m} \sim \mathcal{CN}(0, \frac{1}{1+\eta_{ub_m}})$. Thus, $I_{um}^{\text{Intra}}$ follows exponential distribution with mean $\lambda_{um} \triangleq \frac{\lambda_{u,mb_m}}{1+\eta_{ub_m}}$. The second term $I_{uk}^{\text{Inter}} = \lambda_{u,kb_k}\left|\mathbf{h}_{ub_k}^H\mathbf{w}_{kb_k}\right|^2$ follows exponential distribution with mean $\lambda_{uk} \triangleq \lambda_{u,kb_k}$ due to the independence between $\mathbf{h}_{ub_k}^H$ and $\mathbf{w}_{kb_k}$.

Since $S_u$ follows Gamma distribution with integer parameter $k_u$ and $I_u$ is the sum of exponential distributed RVs, by following the same steps as in Appendix **??**, we can approximate the average data rate for pure CB mode when $\sigma_u^2$ is ignorable as

$$\begin{aligned} \bar{r}_u \approx \sum_{n\neq u}\xi_{un}\Big(&-\ln\zeta_{un} + \sum_{k=1}^{k_u-1}\frac{1}{k} + \frac{(-\zeta_{un})^{k_u}\ln\zeta_{un}}{(1-\zeta_{un})^{k_u}} \\ &- \sum_{k=1}^{k_u-1}\binom{k_u-1}{k}\frac{(1-\rho_{un}^k)(-\zeta_{un})^{k_u-k}}{k(1-\zeta_{un})^{k_u}}\Big), \end{aligned} \tag{22}$$

where $\xi_{un} = \frac{1}{\ln 2}\prod_{n'\neq n}\frac{\lambda_{un}}{\lambda_{un}-\lambda_{un'}}$, $\zeta_{un} = \frac{\lambda_{un}}{\lambda_{u,b_u}\rho_{ub_u}^2}$, and $k_u = N_t - M_{b_u} + 1$.

Then, the average net throughput of the C-RAN system under the backhaul of stringent capacity can be obtained as

$$\bar{R}^*(\mathbf{S}, \mathbf{C}) = \sum_{u=1}^{N_u}(1 - vT)\bar{r}_u^*, \tag{23}$$

where $\bar{r}_u^*$ can be obtained by solving problem (**??**).

*2) Semi-dynamic User-specific Clustering in Pure CB Mode:* Since each UE has only one master BS in pure CB mode, the transmission matrix $\mathbf{S}$ can be determined by letting each UE select the BS with the largest average channel gain as the master BS. Then, the clustering problem reduces to designing the coordination matrix $\mathbf{C}$. We redesign the previously proposed clustering scheme in Section **??** to find $\mathbf{C}$, which is summarized as follows.

1) **Initialization:** Set $\mathbf{S}$ by letting each UE select the BS with the largest average channel gain as the master BS, and let $\mathbf{C}^{(0)} = \mathbf{0}$. Compute $\bar{R}_{\max}^{*(0)} = \bar{R}^*(\mathbf{S}, \mathbf{C}^{(0)})$ and set $i = 1$.
2) **Iteration:**

a) Find the optimal cooperative link maximizing the average net throughput by solving the following problem over $N_u N_f - i$ candidates,

$$\max_{\mathbf{D}^{(i)}} \quad \bar{R}^*(\mathbf{S}, \mathbf{C}^{(i-1)} + \mathbf{D}^{(i)}) \tag{24a}$$

$$s.t. \quad \sum_{u=1}^{N_u} \sum_{b=1}^{N_b} d_{ub}^{(i)} \leq 1, \ d_{ub}^{(i)} \in \{0, 1\} \tag{24b}$$

$$d_{ub}^{(i)} = 0 \ \ (b \notin \mathcal{F}_u) \tag{24c}$$

$$d_{ub}^{(i)} c_{ub}^{(i-1)} = 0, \tag{24d}$$

$$d_{ub}^{(i)} s_{ub} = 0, \tag{24e}$$

$$\sum_{u=1}^{N_u} (c_{ub}^{(i-1)} + s_{ub} + d_{ub}^{(i)}) \leq N_t, \tag{24f}$$

The searching space to solve this problem is $(N_u N_f - i)$, which shrinks with the increase of iteration times $i$.

b) Denote $\mathbf{D}_{\text{opt}}^{(i)}$ as the optimal value of $\mathbf{D}^{(i)}$, and define $\bar{R}_{\max}^{*(i)} \triangleq \bar{R}^*(\mathbf{S}, \mathbf{C}^{(i-1)} + \mathbf{D}_{\text{opt}}^{(i)})$. If $\bar{R}_{\max}^{*(i)} > \bar{R}_{\max}^{*(i-1)}$, update $\mathbf{C}^{(i)} = \mathbf{C}^{(i-1)} + \mathbf{D}_{\text{opt}}^{(i)}$, i.e., add a new cooperative link into the system, set $i = i + 1$, and go back to step 2-a). Otherwise, stop the iteration and $\mathbf{C}^{(i-1)}$ is the final result of the coordination matrix.

The overall computational complexity of this clustering scheme is $\mathcal{O}(\frac{1}{2} N_u^2 N_f^2)$, which is less than the clustering scheme for hybrid CoMP, which is $\mathcal{O}(N_u^2 N_f^2)$.

### *B. Semi-dynamic User-specific Clustering with Pure JT Mode*

The average net throughput in pure JT mode can be obtained from (**??**) by letting $\mathbf{C} = \mathbf{0}$, since each user has no coordinated BS now (all BSs in the cluster are master BSs). Therefore, the user-specific clustering problem reduces to designing the transmission matrix and the clustering scheme designed for the hybrid CoMP can be tailored as follows.

1) **Initialization:** Initialize $\mathbf{S}^{(0)}$ by letting each UE select the BS with the largest average channel gain as a master BS. Compute $\bar{R}_{\max}^{(0)} = \bar{R}(\mathbf{S}^{(0)}, \mathbf{0})$, and set $i = 1$.

2) **Iteration:**

a) Find the optimal transmission link maximizing the average net throughput by solving the following problem over $N_u N_f - i$ candidates,

$$\max_{\mathbf{D}^{(i)}} \quad \bar{R}(\mathbf{S}^{(i-1)} + \mathbf{D}^{(i)}, \mathbf{0}) \tag{25a}$$

$$s.t. \quad \sum_{u=1}^{N_u} \sum_{b=1}^{N_b} d_{ub}^{(i)} \leq 1, \ d_{ub}^{(i)} \in \{0, 1\} \tag{25b}$$

$$d_{ub}^{(i)} = 0 \ \ (b \notin \mathcal{F}_u), \tag{25c}$$

$$d_{ub}^{(i)} s_{ub}^{(i-1)} = 0 \tag{25d}$$

$$\sum_{u=1}^{N_u} (s_{ub}^{(i-1)} + d_{ub}^{(i)}) \leq N_t. \tag{25e}$$

b) Denote $\mathbf{D}_{\text{opt}}^{(i)}$ as the optimal value of $\mathbf{D}^{(i)}$, and define $\bar{R}_{\max}^{(i)} \triangleq \bar{R}(\mathbf{S}^{(i-1)} + \mathbf{D}_{\text{opt}}^{(i)}, \mathbf{0})$. If $\bar{R}_{\max}^{(i)} > \bar{R}_{\max}^{(i-1)}$, update $\mathbf{S}^{(i)} = \mathbf{S}^{(i-1)} + \mathbf{D}_{\text{opt}}^{(i)}$, set $i = i + 1$ and go back to step 2-a). Otherwise, stop the iteration and $\mathbf{S}^{(i-1)}$ is the final result of the transmission matrix.

The overall computational complexity of this clustering scheme is $\mathcal{O}(\frac{1}{2} N_u^2 N_f^2)$, which is less than the one designed for the hybrid CoMP, which is $\mathcal{O}(N_u^2 N_f^2)$.

## V. Numerical and Simulation Results

In this section, we evaluate the performance of the proposed semi-dynamic user-specific clustering schemes via simulations. A cellular network with one tier of $N_b = 7$ cells is considered as shown in Fig. **??**, where each BS is located in the center of a hexagon cell and one UE is uniformly placed in each cell.[4] In order to remove the network boundary effect, we employ the wraparound method in the simulations, with which every cell can be regarded as being surrounded by one tier of cells. Unless otherwise specified, the following parameters from [**?**] are used in all the simulations. The radius of the cell is 250 m. The pathloss is modeled as $35.3 + 37.6 \log_{10}(d_{ub})$ in dB, where $d_{ub}$ is the distance from $\text{BS}_b$ to $\text{UE}_u$ in meters. The average downlink SNR for UEs located at the cell edge is set as 20 dB, whose value depends on the downlink transmit power as well as the cell radius. Considering that the downlink transmit power is usually larger than the uplink transmit power, the equivalent average uplink SNR for UEs located at the cell edge is set as 15 dB, where the uplink training overhead $v = 1\%$ is considered according to LTE specification with a 10 ms training period [**?**]. We also evaluate the performance under different downlink cell-edge SNRs and equivalent uplink cell-edge SNRs later to reflect various configurations of cell radius and transmit power corresponding to different types of BSs.[5] The number of antennas at each BS is four. The measurement set size of each UE is fixed as three, i.e., $N_f = 3$. The average net throughput is obtained by averaging over 500 drops, each of which contains 1000 realizations of i.i.d. Rayleigh small-scale channels.

### *A. Evaluating the Approximations in Deriving the Signal Term*

We first evaluate the accuracy of the Gamma approximation $\hat{S}_u \sim \mathbb{G}(\hat{k}_u, \hat{\theta}_u)$ of the received signal term $S_u$ (taking $\text{UE}_1$ as an example) as well as the round-off Gamma approximation $\tilde{S}_u \sim \mathbb{G}(\tilde{k}_u, \tilde{\theta}_u)$ introduced in Section **??**. Since the distributions of $S_u$, $\hat{S}_u$ and $\tilde{S}_u$ depend on the number of master BSs of $\text{UE}_u$, the distances between $\text{UE}_u$ and its master BSs, and the number of served and coordinated UEs by the master BSs of $\text{UE}_u$, we consider four typical scenarios for comparison, where the number of the master BSs of $\text{UE}_1$ ranges from 1 to 3 with different distances[6] and different numbers of served and coordinated UEs as shown in Table **??**.

[4] The case with multiple users distributed in each cell is also simulated but not shown in the paper since the relationship of the compared schemes is similar to the single-user case.

[5] For example, the transmit power and cell radius of macro BSs and pico BSs are 46 dBm and 250 m, 21 dBm and 60 m, respectively [**?**]. The corresponding downlink cell-edge SNR are 15 dB and 20 dB, respectively.

[6] Since the distributions of $S_u$, $\hat{S}_u$ and $\tilde{S}_u$ depend on the distances between $\text{UE}_u$ and its master BSs, $\text{UE}_1$ is first dropped close to one BS and far away from the other two BSs with (100m, 420m, 420m), and then dropped with similar distances among the three BSs with (250m, 300m, 300m) as shown in Table **??**.

TABLE I
SIMULATION SCENARIOS FOR THE JUSTIFICATION OF GAMMA APPROXIMATIONS

| Scenario | Master BSs | Distance between $\text{UE}_1$ and $\text{BS}_1\sim\text{BS}_3$ | $n(\mathcal{S}_b)+n(\mathcal{I}_b)$ | | |
|---|---|---|---|---|---|
| | | | $\text{BS}_1$ | $\text{BS}_2$ | $\text{BS}_3$ |
| S1 | $\text{BS}_1$ | (100m, 420m, 420m) or (250m, 300m, 300m) | 3 | 0 | 0 |
| S2 | $\text{BS}_1$, $\text{BS}_2$ | | 2 | 1 | 0 |
| S3 | $\text{BS}_1$, $\text{BS}_2$, $\text{BS}_3$ | | 1 | 2 | 3 |
| S4 | $\text{BS}_1$, $\text{BS}_2$, $\text{BS}_3$ | | 1 | 1 | 1 |

Fig. 3. Comparison among the simulation result $S_u$, the Gamma approximation $\hat{S}_u$ and the round-off Gamma approximation $\tilde{S}_u$.

Figure **??** compares the numerical results of the Gamma approximation $\hat{S}_u$ and the round-off Gamma approximation $\tilde{S}_u$ with the simulation result of $S_u$ under the scenarios given in Table **??**. We can see that the PDF of $S_u$ under all considered scenarios with different distances between UE and BSs can be well fitted by the Gamma approximation $\hat{S}_u$ (dashed lines) and the impact of the round-off approximation $\tilde{S}_u$ (dot dashed lines) is negligible.

### *B. Evaluating the Approximations in Deriving the Average Net Throughput*

Fig. 4. The accuracy of the analytical result and the impact of weak interference estimation mechanism.

In Fig. **??**, we compare the simulation results of the net throughput with the analytical result given in (**??**), which assumes high SNR and the knowledge of all large-scale fading gains at the CU. Herein, the locations and the clusters for all UEs are fixed as shown in Fig. **??**. We can see that the analytical results approach the simulation results with a less than $10\%$ gap when the downlink cell-edge SNR $\geq 10$ dB and the equivalent uplink cell-edge SNR $\geq 0$ dB. The analytical results become more accurate when the equivalent uplink cell-edge SNR and downlink cell-edge SNR are higher. We also show the accuracy of the analytical results when only the large-scale fading gains in the measurement set are available at the CU, where we introduced a weak interference estimation mechanism. We can see that without the weak interference estimation mechanism (i.e., simply set the weak interference as zeros) the analytical results become higher than the simulation results because of the underestimated interference power. When the mechanism is employed, the analytical results are very close to the simulation results for all SNRs. This is because the weak interference is well estimated now and the noise power is taken into account.

### *C. Performance Evaluation*

We evaluate the performance of the proposed user-specific clustering schemes by simulating the following schemes. Note that the estimated channels are used for precoding in all the following simulations and also used for clustering in the dynamic schemes.

1) *Non-CoMP system* (with legend "Non-CoMP"): Each UE is only served by the BS with the strongest large-scale channel gain and suffers from ICI from all the other BSs.
2) *Semi-Dynamic Clustering in Pure CB Mode* (with legend "Pure-CB"): Each UE is served by a cluster formed by the proposed scheme given in Section **??** for pure CB.
3) *Semi-Dynamic Clustering in Pure JT Mode* (with legend "Pure-JT"): Each UE is served by a cluster formed by the proposed scheme given in Section **??** for pure JT.
4) *Semi-Dynamic Clustering in Hybrid CoMP Mode* (with legend "Hybrid"): Each UE is served by a cluster formed by the proposed scheme given in Section **??** for hybrid CoMP.
5) *Optimal Semi-Dynamic Clustering in Hybrid CoMP Mode* (with legend "Semi-Dynamic Optimal"): This approach gives the optimal semi-dynamic clustering result for hybrid CoMP by solving problem (**??**) with exhaustive searching.
6) *Dynamic Non-overlapped Clustering with Pure JT Mode* (with legend "Dynamic Non-overlapped"): This approach is proposed in [**?**], where every two BSs form a non-overlapped cluster dynamically in each time slot to jointly serve the users within the cluster using JT. In this approach, the instantaneous CSI from all UEs to all BSs is required, which induces large training overhead.
7) *Dynamic Overlapped Clustering with Pure JT Mode* (with legend "Dynamic Overlapped"): This approach is based on the method proposed in [**?**], where each UE chooses a given number of BSs (set as two in simulations to maximize its net throughput) with the strongest channel gains as its cluster. We apply this method with ZF precoder by limiting the number of users served by each BS not larger than $N_t$. This approach requires instantaneous CSI from each UE to the BSs in its cluster.
8) *Optimal Dynamic Clustering in Hybrid CoMP Mode* (with legend "Dynamic Optimal"): This approach finds the optimal dynamic clustering result that maximizes the instantaneous throughput of the network using ZF precoder and the power allocation given in Section **??** by exhaustive searching, whose performance can be regarded as an upper bound of the clustering method proposed in [**?**]. In this approach, the instantaneous CSI is obtained before the clusters are formed by letting $\text{UE}_u$ send training signals to the BSs in the measurement set $\mathcal{F}_u$, which induces more training overhead than those semi-dynamic clustering schemes.

Fig. 5. Average net throughput vs. backhaul capacity, where the uplink training overhead per UE is $v = 1\%$.

In Fig. **??**, we show the simulation results of the average net throughput versus the backhaul capacity.[7] We can see that the CoMP systems with both Pure-JT and Pure-CB outperform

[7]The capacity of backhaul links in existing practical systems is in a range of 100 Mbps $\sim$ 2 Gbps, which corresponds to $C \sim$ [1 bit/s/Hz, 20 bits/s/Hz] for the bandwidth of 100 MHz [**?**].

the Non-CoMP system when the backhaul capacity is high. When the backhaul capacity is higher than $15$ bits/s/Hz, Pure-JT outperforms Pure-CB because there is enough backhaul capacity for data sharing. When the backhaul capacity is lower than $15$ bits/s/Hz, Pure-JT is inferior to Pure-CB. When the backhaul capacity continues to reduce, Pure-JT becomes inferior to Non-CoMP while Pure-CB still outperforms Non-CoMP, because Pure-CB can avoid ICI within each cluster without any data sharing. The clustering scheme in hybrid CoMP mode outperforms both Pure-JT and Pure-CB, and performs close to the optimal semi-dynamic and optimal dynamic solution. Due to the high training overhead and large inter-cluster interference, dynamic non-overlapped clustering performs even worse than the Non-CoMP system. Although dynamic overlapped clustering can reduce the inter-cluster interference, it still has a large performance loss compared to the proposed semi-dynamic clustering since training overhead is not considered when forming the clusters. We also compare the average iteration times of the proposed methods with the optimal semi-dynamic and optimal dynamic clustering methods under different backhaul capacities as shown in Table **??**. We can see that the complexity of the proposed clustering scheme in hybrid CoMP mode is much lower than those of the optimal semi-dynamic and optimal dynamic clustering schemes with exhaustive searching. When the proposed clustering scheme is applied with Pure-JT or Pure-CB mode, the complexity is further reduced.

TABLE II
AVERAGE ITERATION TIMES UNDER DIFFERENT BACKHAUL CAPACITY

| Algorithms | Backhaul Capacity (bits/s/Hz) | | |
|---|---|---|---|
| | 5 | 12 | 20 |
| Semi-Dyn./Dyn. Optimal | $1.05\times10^{10}$ | $1.05\times10^{10}$ | $1.05\times10^{10}$ |
| Hybrid | 168.3 | 168.6 | 172.7 |
| Pure JT | 89.3 | 89.3 | 89.3 |
| Pure CB | 85.8 | 85.8 | 85.8 |

Fig. 6. Average net throughput vs. training overhead per UE $v$, where the backhaul capacity is $C = 15$ bits/s/Hz.

In Fig. **??**, we show the simulation results of the average net throughput versus the training overhead per UE. We can see that the semi-dynamic clustering scheme in hybrid CoMP mode outperforms Pure-JT and Pure-CB, and the optimal dynamic clustering performs no better than the semi-dynamic clustering scheme in hybrid CoMP mode. When the training overhead is low, the performance of all the approaches increases with the training overhead since more training overhead leads to higher equivalent uplink SNR and thus more accurate CSI, which improves the performance especially for CoMP schemes. When the training overhead becomes higher, the performance decreases since the uplink training wastes too much resources, which counteracts the gain of accurate CSI. Moreover, the performance of the three dynamic clustering schemes, including dynamic non-overlapped, dynamic overlapped and optimal dynamic clustering, degrades sharply and becomes inferior to semi-dynamic clustering and even Non-CoMP. This is because the cluster size of dynamic clustering does not depend on the training overhead, while the proposed semi-dynamic clustering schemes can decrease the cluster size of each UE so as to reduce the overall training overhead. Since the cluster size of each UE is small when the training overhead is high, the backhaul load is not heavy, and thus Pure-JT outperforms Pure-CB and performs close to the hybrid CoMP mode.

(a)(b)

Fig. 7. Average net throughput vs. (a) the size of measurement set $N_f$ per UE and (b) the inverse of measurement set threshold $1/\beta$ in dB, where the backhaul capacity is $C = 15$ bits/s/Hz.

In Fig. **??**, we show the simulation results of the average net throughput versus the size of measurement set, where both the fixed-size measurement set and the threshold-based measurement set selection methods in Section **??**-D are considered. We can see that all the CoMP systems with the proposed semi-dynamic clustering schemes in Pure-JT, Pure-CB and hybrid mode outperform the Non-CoMP system when the measurement set is larger than one or when $1/\beta > 0$ dB, and reach the maximum throughput when $N_f = 3$ or $1/\beta = 15$ dB. This verifies that a UE will gain little from choosing faraway BSs as its master BSs or coordinated BSs. Therefore, in practical systems when implementing the proposed clustering schemes, the measurement set should be small, which can reduce the computational complexity that is proportional to $N_f^2$ as we showed before.

Fig. 8. Average net throughput vs. downlink cell-edge SNR, where the backhaul capacity is $C = 15$ bits/s/Hz.

In Fig. **??**, we show the simulation results of the average net throughput versus downlink cell-edge SNR. It is shown that the semi-dynamic clustering scheme in hybrid CoMP mode outperforms Pure-JT and Pure-CB for high downlink cell-edge SNR, because in this case the network is backhaul capacity limited. When the downlink cell-edge SNR is low, the gap between Pure-JT and Pure-CB becomes larger and the performance of Pure-JT is close to the hybrid mode, because the backhaul capacity is not limited now and the hybrid mode reduces to Pure-JT mode.

Fig. 9. Average net throughput vs. equivalent uplink cell-edge SNR, where the backhaul capacity is $C = 15$ bits/s/Hz.

Finally, in Fig. **??** we show the simulation results of the average net throughput versus equivalent uplink cell-edge SNR, which affects the accuracy of channel estimation, and we also give the performance with perfect CSI for comparison. We can see that when the equivalent uplink SNR continues to increase, the performance of all the methods approach to the performance with perfect CSI. For high equivalent uplink cell-edge SNR, all the CoMP systems with the proposed semi-dynamic clustering schemes in Pure-JT, Pure-CB and hybrid mode outperform the Non-CoMP system, since in this case the CSI for CoMP is accurate. When the equivalent uplink cell-edge SNR is larger than 20 dB, Pure-CB outperforms

Pure-JT and the gap between the hybrid mode and Pure-JT is large, because the downlink data rate is high now and the performance of the network is backhaul capacity limited. When the equivalent uplink cell-edge SNR is low, the CSI is not accurate enough for both Pure-JT and Pure-CB. Then, all the CoMP systems in Pure-JT, Pure-CB and hybrid mode degrade to the Non-CoMP system, where no cooperative link or extra transmission link is added to the Non-CoMP system with the proposed semi-dynamic clustering schemes.

## VI. Conclusions

In this paper, we studied semi-dynamic user-specific clustering for downlink C-RAN with CoMP transmission under limited-capacity backhaul. Considering that backhaul links with different capacities may be employed in C-RAN, we introduced a hybrid CoMP transmission mode. Under the hybrid mode, we derived the closed-form expression of the asymptotical average data rate of C-RAN system in high SNR regime, and introduced a weak interference estimation mechanism to improve the accuracy of the asymptotical results in general SNRs. By taking into account the training overhead for channel estimation in C-RAN, we designed the semi-dynamic user-specific clustering scheme, aimed at maximizing the average net throughput of the C-RAN system, subject to the constraint on the backhaul capacity. The proposed clustering scheme only depends on large-scale channel gains and therefore can be operated in a semi-dynamic manner. Moreover, the scheme is of low complexity and performs close to the optimal solution found by exhaustive searching. Then, we tailored the proposed clustering scheme to two special cases respectively with very stringent and unlimited backhaul capacity in order to further reduce complexity, where the hybrid CoMP degenerates to the pure CB and pure JT. Simulations validated the analytical results, and showed that the proposed semi-dynamic user-specific clustering schemes are superior to the Non-CoMP systems, and outperform the dynamic clustering schemes when the training overhead is large.

## Appendix A
## Summary of the algorithm for computing $T$

The algorithm proposed in [**?**] for computing $T$ is based on graph theory. We take the clustering result shown in Fig. **??** as an example, whose graph representation can be shown in Fig. **??**, where each vertex of the graph represents a UE and each edge in the graph represents two UE sharing no less than one BS in their clusters. For instance, $\text{BS}_1$ is in the clusters of both $\text{UE}_1$ and $\text{UE}_4$ as shown in Fig. **??**, so there is an edge between $\text{UE}_1$ and $\text{UE}_4$. The principle for allocating orthogonal training resources among the UEs is to use the minimal number of total training resources to ensure that any two connected UEs have orthogonal training resources. The procedure for computing $T$ in Fig. **??** can be briefly summarized as follows.

1) Compute the degree of each vertex, which is $4, 2, 2, 3, 1, 2, 2$ for $\text{UE}_1 \sim \text{UE}_7$.
2) Sort the vertices in a descending order by their degrees.
3) Allocate the orthogonal training resources for each vertex sequentially to ensure that any two connected UEs employ orthogonal training resources and the total number of used training resources is minimized.

The allocation results are shown in Fig. **??**, where different colors represent orthogonal training resources and the total number of used orthogonal training resources is 3, i.e., $T = 3$.

Fig. 10. Graph representation of Fig. **??** and the training resources allocation result.

## Appendix B
## Proof of Proposition 1

The signal term can be approximated as

$$\begin{aligned} S_u &= \Big| \sum_{b\in\mathcal{T}_u} \sqrt{\lambda_{u,b}} \mathbf{h}_{ub}^H \mathbf{w}_{ub} \Big|^2 \\ &= \Big| \sum_{b\in\mathcal{T}_u} \sqrt{\lambda_{u,b}} (\rho_{ub} \hat{\mathbf{h}}_{ub} + \mathbf{e}_{ub})^H \mathbf{w}_{ub} \Big|^2 \\ &\overset{(a)}{\approx} \Big| \sum_{b\in\mathcal{T}_u} \sqrt{\lambda_{u,b}} \rho_{ub} \hat{\mathbf{h}}_{ub}^H \mathbf{w}_{ub} \Big|^2 \\ &\overset{(b)}{=} \Big( \sum_{b\in\mathcal{T}_u} \sqrt{\lambda_{u,b}} \rho_{ub} \hat{\mathbf{h}}_{ub}^H \mathbf{w}_{ub} \Big)^2 \triangleq \Big( \sum_{b\in\mathcal{T}_u} S_{u,b} \Big)^2, \end{aligned} \tag{26}$$

where approximation (a) ignores the term $\mathbf{e}_{ub}$, which is accurate when the uplink SNR $\eta_{ub}$ is high as verified through simulations, and step (b) turns $|\cdot|$ into $(\cdot)$ since $\hat{\mathbf{h}}_{ub}^H \mathbf{w}_{ub}$ is a positive real number with ZF precoding. Since $|\hat{\mathbf{h}}_{ub}^H \mathbf{w}_{ub}|^2$ follows Gamma distribution $\mathbb{G}(K_b, 1)$ with $K_b = N_t - M_b + 1$ [**?**] and $\lambda_{u,b} = \alpha_{ub}^2 \frac{\alpha_{ub}^2 P}{\sum_{k\in\mathcal{S}_b} \alpha_{kb}^2}$ is independent from the small scale channels, we can obtain that $S_{u,b}^2 = \lambda_{u,b} \rho_{ub}^2 |\hat{\mathbf{h}}_{ub}^H \mathbf{w}_{ub}|^2$ follows Gamma distribution $\mathbb{G}(K_b, \lambda_{u,b}\rho_{ub}^2)$ and its square root $S_{u,b}$ follows Nakagami-$m$ distribution with parameters $\Omega_{ub} = K_b \lambda_{u,b} \rho_{ub}^2$ and $m_b = K_b$.

## Appendix C
## Proof of Proposition 2

Denote $T_u$ as the size of $\mathcal{T}_u$, where $\mathcal{T}_u = \{b_{u1}, b_{u2}, \cdots, b_{uT_u}\}$. By recursive use of the binomial theorem [**?**], the $n$th moment of $S_u$ can be found as

$$\begin{aligned} &\mathbb{E}\{S_u^n\} \overset{(a)}{\approx} \mathbb{E}\Big\{ \Big( \sum_{b\in\mathcal{T}_u} S_{u,b} \Big)^{2n} \Big\} \\ &= \sum_{n_1=0}^{2n} \sum_{n_2=0}^{n_1} \cdots \sum_{n_{T_u-1}}^{n_{T_u-2}} \binom{n}{n_1} \binom{n_1}{n_2} \cdots \binom{n_{T_u-2}}{n_{T_u-1}} \mathbb{E}\big\{ S_{u,b_1}^{2n-n_1} \big\} \\ &\quad \times \mathbb{E}\big\{ S_{u,b_2}^{n_1-n_2} \big\} \cdots \mathbb{E}\big\{ S_{u,b_{T_u-1}}^{n_{T_u-2}-n_{T_u-1}} \big\} \mathbb{E}\big\{ S_{u,b_{T_u}}^{n_{T_u-1}} \big\}, \end{aligned} \tag{27}$$

where step (a) comes from (**??**) and the $n$th moment of $S_{u,b}$ is given by [**?**]

$$\mathbb{E}\{S_{u,b}^n\} = \frac{\Gamma(m_b + \frac{1}{2})}{\Gamma(m_b)} \left( \frac{\Omega_{ub}}{m_b} \right)^{\frac{n}{2}}. \tag{28}$$

Then we can match the first two moments of $\hat{S}_u$ and $S_u$ as

$$\mathbb{E}\{S_u\} = \mathbb{E}\{\hat{S}_u\} = \hat{k}_u\hat{\theta}_u, \tag{29a}$$
$$\mathbb{E}\{S_u^2\} = \mathbb{E}\{\hat{S}_u^2\} = \hat{k}_u\hat{\theta}_u^2 + \hat{k}_u^2\hat{\theta}_u^2. \tag{29b}$$

By solving (**??**), $\hat{k}_u$ and $\hat{\theta}_u$ can be found as

$$\hat{\theta}_u = \frac{\mathbb{E}\{S_u^2\} - \mathbb{E}^2\{S_u\}}{\mathbb{E}\{S_u\}}, \tag{30a}$$
$$\hat{k}_u = \frac{\mathbb{E}^2\{S_u\}}{\mathbb{E}\{S_u^2\} - \mathbb{E}^2\{S_u\}}. \tag{30b}$$

## Appendix D
## Proof of Proposition 3

Since $\tilde{S}_u$ is independent from $I_u$, with (**??**), we first take the expectation over $I_u$ for $\sigma_u^2 \to 0$ and obtain [**?**, eq.(28)]

$$\bar{r}_u = \mathbb{E}_{S_u,I_u}\left\{\log_2(1+\gamma_u)\right\} \approx \mathbb{E}_{\tilde{S}_u,I_u}\left\{\log_2\left(1+\tfrac{\tilde{S}_u}{I_u}\right)\right\}$$
$$= \sum_{m\neq u}\frac{\delta_{um}\lambda_{um}}{\ln 2}\mathbb{E}_{\tilde{S}_u}\left\{\varepsilon + \ln\tfrac{\tilde{S}_u}{\lambda_{um}} - e^{\frac{\tilde{S}_u}{\lambda_{um}}}\mathrm{Ei}\left(-\tfrac{\tilde{S}_u}{\lambda_{um}}\right)\right\}, \tag{31}$$

where $\mathrm{Ei}(x) = -\int_{-x}^{\infty}\frac{e^{-t}}{t}dt$ denotes the exponential integral and $\varepsilon$ is the Euler-Mascheroni constant. Then, by taking the expectation over $\tilde{S}_u$, we have

$$\mathbb{E}_{\tilde{S}_u}\left\{\varepsilon + \ln\tfrac{\tilde{S}_u}{\lambda_{um}} - e^{\frac{\tilde{S}_u}{\lambda_{um}}}\mathrm{Ei}\left(-\tfrac{\tilde{S}_u}{\lambda_{um}}\right)\right\}$$
$$=\varepsilon + \int_0^{\infty}\left(\ln\tfrac{s}{\lambda_{um}} - e^{\frac{s}{\lambda_{um}}}\mathrm{Ei}\big(-\tfrac{s}{\lambda_{um}}\big)\right)\tfrac{s^{\tilde{k}_u-1}e^{-\frac{s}{\tilde{\theta}_u}}}{\tilde{\theta}_u^{\tilde{k}_u}(\tilde{k}_u-1)!}ds$$
$$=\varepsilon + \int_0^{\infty}\left(\tfrac{x^{\tilde{k}_u-1}e^{-x}}{(\tilde{k}_u-1)!}\ln\tfrac{x}{\tilde{\zeta}_{um}} - e^{\frac{x}{\tilde{\zeta}_{um}}}\mathrm{Ei}\big(-\tfrac{x}{\tilde{\zeta}_{um}}\big)\tfrac{x^{\tilde{k}_u-1}e^{-x}}{(\tilde{k}_u-1)!}\right)dx, \tag{32}$$

where $\tilde{\zeta}_{um} = \frac{\lambda_{um}}{\tilde{\theta}_u}$. The integral in (**??**) can be derived by [**?**, p. 569, eq. (4.352.1)] and [**?**, p. 633, eq. (6.228.2)] as

$$\Psi(\tilde{k}_u) - \ln\tilde{\zeta}_{um} + \tfrac{1}{\tilde{k}_u}\,{}_2F_1(1,\tilde{k}_u;\tilde{k}_u+1;1-\tfrac{1}{\tilde{\zeta}_{um}}), \tag{33}$$

where $\Psi(x)$ and ${}_2F_1(a,b;c;x)$ are the Digamma function and the Gauss Hypergeometric function, respectively.

Noting that $\tilde{S}_u \sim \mathbb{G}(\tilde{k}_u, \tilde{\theta}_u)$ is the round-off approximation of Gamma distribution, where $\tilde{k}_u$ is an integer, $\Psi(\tilde{k}_u)$ can be expressed in closed form as [**?**, p. 894, eq. (8.365.4)]

$$\Psi(\tilde{k}_u) = -\varepsilon + \sum_{k=1}^{\tilde{k}_u-1}\frac{1}{k}, \tag{34}$$

and

$$\tfrac{1}{\tilde{k}_u}\,{}_2F_1(1,\tilde{k}_u;\tilde{k}_u+1;1-\tfrac{1}{\tilde{\zeta}_{um}})$$
$$\overset{(a)}{=}\tfrac{\tilde{\zeta}_{um}}{\tilde{k}_u}\,{}_2F_1(1,1;\tilde{k}_u+1,1-\tilde{\zeta}_{um})$$
$$\overset{(b)}{=}\tilde{\zeta}_{um}\int_0^1\tfrac{(1-t)^{\tilde{k}_u-1}}{1-(1-\tilde{\zeta}_{um})t}dt$$
$$=\tfrac{\tilde{\zeta}_{um}}{(1-\tilde{\zeta}_{um})^{\tilde{k}_u-1}}\int_0^1\tfrac{\left(1-(1-\tilde{\zeta}_{um})t-\tilde{\zeta}_{um}\right)^{\tilde{k}_u-1}}{1-(1-\tilde{\zeta}_{um})t}dt$$
$$=\tfrac{\tilde{\zeta}_{um}}{(1-\tilde{\zeta}_{um})^{\tilde{k}_u-1}}\times$$
$$\sum_{k=0}^{\tilde{k}_u-1}\tbinom{\tilde{k}_u-1}{k}(-\tilde{\zeta}_{um})^{\tilde{k}_u-k-1}\int_0^1\left(1-(1-\tilde{\zeta}_{um})t\right)^{k-1}dt$$
$$=\tfrac{(-\tilde{\zeta}_{um})^{\tilde{k}_u}\ln\tilde{\zeta}_{um}}{(1-\tilde{\zeta}_{um})^{\tilde{k}_u}} - \sum_{k=1}^{\tilde{k}_u-1}\tbinom{\tilde{k}_u-1}{k}\tfrac{(1-\tilde{\zeta}_{um}^k)(-\tilde{\zeta}_{um})^{\tilde{k}_u-k}}{k(1-\tilde{\zeta}_{um})^{\tilde{k}_u}}, \tag{35}$$

where step (a) follows the transformation by [**?**, p. 998, eq. (9.131.1)] and step (b) comes from the integral representation of the Gauss Hypergeometric function in [**?**, p. 995, eq. (9.111)].

Finally, by substituting (**??**) and (**??**) into (**??**), defining $\xi_{um} \triangleq \frac{\lambda_{um}\delta_{um}}{\ln 2}$, and further considering (**??**), Proposition 3 is proved.