Underlaid Spectrum Sharing for Cell-Free Massive MIMO-NOMA

Fatemeh Rezaei, *Student Member, IEEE*, Ali Reza Heidarpour, *Student Memeber, IEEE*, Chintha Tellambura, *Fellow, IEEE*, and Aliakbar Tadaion, *Senior Member, IEEE*

***Abstract*— We investigate the achievable rate of a hybrid cell-free (CF) massive multiple-input multiple-output (MIMO) non-orthogonal multiple-access (NOMA) underlaid below a primary massive MIMO. We propose a low complexity sub-optimal user-clustering method and derive the closed-form sum rate expression for Rayleigh fading channels by considering the effects of intra-cluster pilot contamination, inter-cluster interference, imperfect successive interference cancellation (SIC) and statistical downlink channel state information (CSI) at secondary users (SUs). Our results reveal that NOMA based underlay CF massive MIMO-NOMA can exploit the scarce spectrum bands more efficiently than its counterpart orthogonal multiple-access (OMA).**

***Index Terms*— NOMA, cell-free massive MIMO, underlay spectrum sharing.**

I. Introduction

THE hybrid of cell-free (CF) massive multi-input multiple-output (MIMO) and non-orthogonal multiple access (NOMA) yields significant spectral efficiency (SE) gains. CF massive MIMO uses a large number of spatially distributed access points (APs) to serve many users without cell boundaries (cell-free) [1]. NOMA serves multiple users simultaneously by exploiting channel gain disparities and applying successive interference cancellation (SIC) based detection [2]–[4]. To further enhance SE, these two technologies can be integrated with underlay spectrum sharing. In underlay spectrum sharing, primary users (PUs) and secondary users (SUs) access the same spectrum simultaneously. The transmit power of SUs must be curtailed to ensure that any potential harm to the primary network is minimized [5]–[7].

Motivation and Our Contributions: The hybrid of primary massive MIMO and underlaid CF massive MIMO-NOMA has not yet been investigated. That is the main problem addressed in this letter. There are several reasons why this configuration is of interest. First, since massive MIMO is widely deployed for fifth generation (5G) cellular, it can be exploited effectively as a primary BS in underlay cognitive settings. Second, interference management is easier because distributed secondary transmissions may reduce overall interference due to shorter end-to-end secondary links. Moreover, APs could be easily turned off/on adapting to the network condition; which improves the performance of PU/SU links, an important consideration in underlay cognitive settings. Third, distributed APs can also provide better coverage and increase the macro diversity gain for the secondary network. Finally, massive MIMO-OMA (orthogonal multiple access) capacity faces two fundamental roadblocks: (1) to achieve best SE, the number of antennas (equivalently the number of radio frequency chains) must exceed the number of users, which is not true for **dense networks** [8], and (2) the **channel coherence interval** also limits the capacity [3]. Therefore, rather than adding more hardware to handle more users, NOMA offers a software solution to these two physical challenges.

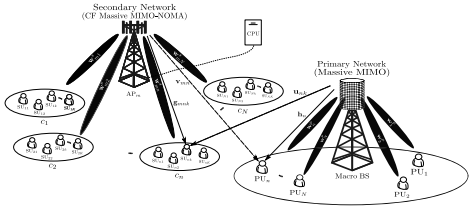

Fig. 1. System model of a hybrid CF massive MIMO-NOMA underlaid below a primary massive MIMO.

Due to the aforementioned reasons, this system model in Fig. 1 is of practical importance. However, in order to take advantage of power domain multiplexing for the best performance of NOMA, the users should have maximum channel gain differences. This necessitates that the users should be clustered so that the users in each cluster have the channel gain difference as large as possible. Several user clustering algorithms have been investigated in the context of cell-centric (non-CF) NOMA systems [9]. However, none of these algorithms can be applied directly to CF-NOMA. In this letter, we thus propose a low complexity sub-optimal user-clustering method that significantly improves the achievable sum rate of the CF massive MIMO-NOMA network.

The main contributions of this work can be summarized as follows: (1) We investigate the performance of a secondary (cognitive) CF massive MIMO-NOMA system underlaid a primary massive MIMO system. The primary macro BS and the secondary multi-antenna APs employ maximum ratio transmission (MRT) beamforming. We derive the closed-form secondary downlink sum rate by considering the effects of intra-cluster pilot contamination, inter-cluster interference, imperfect SIC and statistical downlink channel state information (CSI) at SUs. (2) We propose a sub-optimal user-clustering method which uses Jaccard distance coefficient (a.k.a. Tanimoto) [10] to find the most dissimilar SUs in the secondary network.

II. System Model and Preliminaries

A. System and Channel Models

In our system, CF massive MIMO-NOMA is the secondary network and massive MIMO is the primary network (Fig. 1).

Manuscript received December 7, 2019; accepted December 29, 2019. Date of publication January 13, 2020; date of current version April 9, 2020. The associate editor coordinating the review of this letter and approving it for publication was L. Dai. *(Corresponding author: Fatemeh Rezaei.)*

Fatemeh Rezaei is with the Department of Electrical and Computer Engineering, University of Alberta, Edmonton, AB T6G 2R3, Canada, and also with the Department of Electrical Engineering, Yazd University, Yazd 89195-741, Iran (e-mail: rezaeidi@ualberta.ca).

Ali Reza Heidarpour and Chintha Tellambura are with the Department of Electrical and Computer Engineering, University of Alberta, Edmonton, AB T6G 2R3, Canada (e-mail: heidarpo@ualberta.ca; ct4@ualberta.ca).

Aliakbar Tadaion is with the Department of Electrical Engineering, Yazd University, Yazd 89195-741, Iran (e-mail: tadaion@yazd.ac.ir).

Digital Object Identifier 10.1109/LCOMM.2020.2966195

Both the networks utilize time-division duplexing (TDD) and are synchronized perfectly to prevent the PU-SU interference [5]. The primary macro BS equipped with M^p antennas serve N single-antenna PUs, simultaneously. In the secondary network, M^s APs, each equipped with L antennas, jointly serve KN single-antenna SUs. The SUs are grouped into N clusters with K $(K \geq 2)$ users per cluster and NOMA is applied among the SUs in the same cluster. All the secondary APs are connected to a CPU via an error-free fronthaul network to achieve coherent processing. Secondary APs and the CPU exchange only payload data and power control coefficients that change slowly [5].

In the primary network, the channel between the nth PU and the macro BS is $\mathbf{h}_n \in \mathcal{C}^{M^p \times 1}$. For the secondary network, $\mathbf{g}_{mnk} \in \mathcal{C}^{L \times 1}$ represents the channel between the kth SU in the nth cluster and the mth secondary AP. Besides, the channel between the kth SU in the nth cluster and the macro BS is $\mathbf{u}_{nk} \in \mathcal{C}^{M^p \times 1}$. Moreover, $\mathbf{v}_{mn} \in \mathcal{C}^{L \times 1}$ is the channel between the mth secondary AP and the nth PU. A unified representation of all four channels is $\mathbf{a} = {\beta_{\mathbf{a}}}^{1/2}\bar{\mathbf{a}}$, where $\mathbf{a} \in \{\mathbf{h}_n, \mathbf{g}_{mnk}, \mathbf{u}_{nk}, \mathbf{v}_{mn}\}$ and $\beta_{\mathbf{a}}$ accounts for the large-scale pathloss and shadowing. Whereas $\bar{\mathbf{a}} \sim \mathcal{CN}\left(\mathbf{0}, \mathbf{I}_{M^P|L}\right)$ captures Rayleigh fading.

B. Uplink Pilot Transmission

In TDD, by exploiting channel reciprocity, the primary macro BS and the secondary APs locally estimate the channels via the uplink pilot sequences transmitted by the PUs and SUs. The N PUs are assigned with orthogonal pilot sequences of length $\tau \geq N$ samples [1]. To minimize channel estimation overhead, the same pilots are reused by the secondary network and one pilot sequence is assigned per cluster. The pilot sequence for the kth SU in the nth cluster and the nth PU is $\sqrt{\tau}\boldsymbol{\varphi}_n \in \mathcal{C}^{\tau \times 1}$ satisfying $\|\boldsymbol{\varphi}_n\|^2 = 1$.

As the primary and secondary users transmit the pilot sequences in the uplink, the mth secondary AP estimates $\mathbf{g}_{mnk}$ using minimum mean square error (MMSE) estimation [11]. The MMSE estimate of $\mathbf{g}_{mnk}$ can be expressed as $\hat{\mathbf{g}}_{mnk} = c_{mnk}\tilde{\mathbf{y}}^s_{mn}$, where c_{mnk} is given by [11]

$$c_{mnk} = \sqrt{\tau p_p}\beta_{\mathbf{g}_{mnk}}\left(1+\tau p_p\left(\sum_{i=1}^{K}\beta_{\mathbf{g}_{mni}}+\beta_{\mathbf{v}_{mn}}\right)\right)^{-1}, \quad (1)$$

and $\tilde{\mathbf{y}}^s_{mn}$, the projected received pilot signal at the mth secondary AP onto $\boldsymbol{\varphi}_n$, is given as

$$\tilde{\mathbf{y}}^s_{mn} = \sqrt{\tau p_p}\sum_{k=1}^{K}\mathbf{g}_{mnk} + \sqrt{\tau p_p}\mathbf{v}_{mn} + \tilde{\mathbf{n}}^s_{mn}, \quad (2)$$

where p_p is the pilot transmit power and $\tilde{\mathbf{n}}^s_{mn} \sim \mathcal{CN}(\mathbf{0}, \mathbf{I}_L)$.

Since $\tilde{\mathbf{y}}^s_{mn}$ is Gaussian distributed, $\hat{\mathbf{g}}_{mnk}$ can be written as, $\hat{\mathbf{g}}_{mnk} = \sqrt{\theta_{\hat{\mathbf{g}}_{mnk}}}\nu^s_{mn}$, where $\nu_{mn} \sim \mathcal{CN}(\mathbf{0}, \mathbf{I}_L)$ and $\theta_{\hat{\mathbf{g}}_{mnk}} = \frac{1}{L}\mathbb{E}\left\{\|\hat{\mathbf{g}}_{mnk}\|^2\right\}$ is equal to

$$\theta_{\hat{\mathbf{g}}_{mnk}} = \tau p_p \beta^2_{\mathbf{g}_{mnk}}\left(1+\tau p_p\left(\sum_{i=1}^{K}\beta_{\mathbf{g}_{mni}}+\beta_{\mathbf{v}_{mn}}\right)\right)^{-1}. \quad (3)$$

The channel estimation error can then be defined as $\epsilon_{\mathbf{g}_{mnk}} = \mathbf{g}_{mnk} - \hat{\mathbf{g}}_{mnk}$ where $\epsilon_{\mathbf{g}_{mnk}} \sim \mathcal{CN}(\mathbf{0}, (\beta_{\mathbf{g}_{mnk}} - \theta_{\hat{\mathbf{g}}_{mnk}})\mathbf{I}_L)$. Moreover, since the SUs in the same cluster use the same pilot sequence, their channels are parallel; i.e., we have $\hat{\mathbf{g}}_{mnk} = \left(\beta_{\mathbf{g}_{mnk}}/\beta_{\mathbf{g}_{mni}}\right)\hat{\mathbf{g}}_{mni}$. Similarly, for the primary network, the estimated channel is $\hat{\mathbf{h}}_n \sim \mathcal{CN}(\mathbf{0}, \theta_{\hat{\mathbf{h}}_n}\mathbf{I}_{M^p})$ [11], where

$$\theta_{\hat{\mathbf{h}}_n} = \tau p_p \beta^2_{\hat{\mathbf{h}}_n}\left(1+\tau p_p\left(\sum_{i=1}^{K}\beta_{\mathbf{u}_{ni}}+\beta_{\mathbf{h}_n}\right)\right)^{-1}. \quad (4)$$

C. Downlink Data Transmission Model

For the secondary network, the superposition coded data signal for the K SUs in the nth cluster is expressed as [4]

$$s^s_n = \sqrt{p^s_t}\sum_{k=1}^{K}\sqrt{\lambda^s_{nk}}s^s_{nk}, \quad n = 1, \ldots, N, \quad (5)$$

where s^s_{nk}, λ^s_{nk} and p^s_t denote the data signal, power allocation coefficient for the kth SU in the nth cluster and the total transmitted power by each secondary AP. The set of power coefficients satisfies $\sum_{n=1}^{N}\sum_{k=1}^{K}\lambda^s_{nk} = 1$. Furthermore, the different data signals are mutually uncorrelated; $\mathbb{E}\{s^s_{nk}(s^s_{mi})^*\} = \delta(m-n)\delta(k-i)$, where $m, n \in \{1, 2, \ldots, N\}$ and $k, i \in \{1, 2, \ldots, K\}$. Therefore, $\mathbb{E}\{|s^s_n|^2\} = p^s_t\sum_{k=1}^{K}\lambda^s_{nk} = p^s_t\lambda_n$, where $\lambda^s_n = \sum_{k=1}^{K}\lambda^s_{nk}$ accounts for the power allocation coefficient for the nth cluster.

The mth secondary AP $(\forall m)$ transmits the signal $\mathbf{x}^s_m = \sum_{n=1}^{N}\mathbf{w}^s_{mn}s^s_n$, where $\mathbf{w}^s_{mn}$ is the spatial directivity of the signals sent to the SUs in the nth cluster by the mth secondary AP. Note that each secondary AP precodes the transmitted signals for all the SUs in the same cluster with the same beamforming vector $\mathbf{w}^s_{mn}$, i.e., each secondary AP has N precoding vectors. Similarly, the primary macro BS transmits the signal $\mathbf{x}^p_m = \sqrt{p^p_t}\sum_{n=1}^{N}\mathbf{w}^p_n\sqrt{\lambda^p_n}s^p_n$, where p^p_t, λ^p_n and $\mathbf{w}^p_n$ are the total transmit power, the power coefficients for the PUs ($\sum_{n=1}^{N}\lambda^p_n = 1$), and the precoding vector for the nth PU [4].

We employ MRT beamforming to precode data signals at the primary macro BS and the secondary APs. Therefore, the precoding vectors are given as [1] $\mathbf{w}^s_{mn} = \hat{\mathbf{g}}_{mnk}/\sqrt{\mathbb{E}\{\|\hat{\mathbf{g}}_{mnk}\|^2\}}$ and $\mathbf{w}^p_n = \hat{\mathbf{h}}_n/\sqrt{\mathbb{E}\{\|\hat{\mathbf{h}}_n\|^2\}}$. In underlay spectrum sharing, the transmit power of the secondary APs are constrained to manage the interference inflicted at the PUs [7]

$$p^s_{\mathrm{t}} = \min\left(p^s_{t,\max}, I_{p_1}/P(z_1), \ldots, I_{p_N}/P(z_N)\right), \quad (6)$$

where $p^s_{t,\max}$ is the maximum transmit power by each secondary AP and I_{p_n} is the interference temperature (maximum tolerable interference level) for the nth PU [7]. The interference power inflicted at the nth PU by the secondary network is scaled by $P(z_n)$ in which z_n is

$$z_n = \sum_{m=1}^{M^s}\sum_{n'=1}^{N}\lambda^s_{n'}\mathbf{v}^{\mathrm{H}}_{mn}\mathbf{w}_{mn'}. \quad (7)$$

Regarding (7), $P(z_n)$ can be derived as follows (Appendix A)

$$P(z_n) = \sum_{n'=1}^{N}\lambda^s_{n'}\sum_{m=1}^{M^s}\beta_{\mathbf{v}_{mn}} + \lambda^s_n\left(\sum_{m=1}^{M^s}\frac{\sqrt{L\tau p_p}}{\sqrt{\zeta^s_{mn}}}\beta_{\mathbf{v}_{mn}}\right)^2, \quad (8)$$

where $\zeta^s_{mn} \triangleq 1 + \tau p_p(\sum_{i=1}^{K}\beta_{\mathbf{g}_{mni}} + \beta_{\mathbf{v}_{mn}})$.

To apply power domain NOMA, we assume that the SUs in the nth cluster are ordered based on their effective channel strength as $\Gamma_{n1} \geq \Gamma_{n2} \geq \cdots \geq \Gamma_{nK}$, where $\Gamma_{nk} = \mathbb{E}\{|\sum_{m=1}^{M^s} \hat{\mathbf{g}}_{mnk}^{\mathrm{H}} \mathbf{w}_{mn}^s|^2\}$.

Higher powers are allocated to the SUs with lower channel strength; i.e., $\lambda_{n1}^s \leq \lambda_{n2}^s \leq \ldots \leq \lambda_{nK}^s$. Hence, the kth SU applies SIC to decode its own signal. More precisely, it decodes the signals of the SUs with higher powers and treats the others as interference. In particular, the kth SU decodes the signal intended for the ith SU ($\forall i \geq k$) and treats the signals of the other SUs ($\forall i < k$) as interference.

Since the KN SUs are served simultaneously by M^s secondary APs in the presence of the primary network, using the statistical CSI knowledge of the effective channels at SUs i.e., $\mathbb{E}\left\{\mathbf{g}_{mnk}^{\mathrm{H}} \mathbf{w}_{mn}^s\right\}$ ($\forall m, n, k$), the received signal at the kth SU in the nth cluster can be expressed as

$$
\begin{aligned}
y_{nk}^s = & \underbrace{\sqrt{p_t^s} \sum_{m=1}^{M^s} \sqrt{\lambda_{nk}^s} \mathbb{E}\left\{\mathbf{g}_{mnk}^{\mathrm{H}} \mathbf{w}_{mn}^s\right\} s_{nk}^s}_{\text{desired signal}} \\
& + \underbrace{\sqrt{p_t^s} \sum_{m=1}^{M^s} \sqrt{\lambda_{nk}^s} \left(\mathbf{g}_{mnk}^{\mathrm{H}} \mathbf{w}_{mn}^s - \mathbb{E}\left\{\mathbf{g}_{mnk}^{\mathrm{H}} \mathbf{w}_{mn}^s\right\}\right) s_{nk}^s}_{\text{beamforming gain uncertainty}} \\
& + \underbrace{\sqrt{p_t^s} \sum_{m=1}^{M^s} \sum_{i=1}^{k-1} \sqrt{\lambda_{ni}^s} \mathbf{g}_{mnk}^{\mathrm{H}} \mathbf{w}_{mn}^s s_{ni}^s}_{\text{intra-cluster interference after SIC}} \\
& + \underbrace{\sqrt{p_t^s} \sum_{m=1}^{M^s} \sum_{i=k+1}^{K} \sqrt{\lambda_{ni}^s} \left(\mathbf{g}_{mnk}^{\mathrm{H}} \mathbf{w}_{mn}^s s_{ni}^s - \mathbb{E}\left\{\mathbf{g}_{mnk}^{\mathrm{H}} \mathbf{w}_{mn}^s\right\} \hat{s}_{ni}^s\right)}_{\text{residual interference due to imperfect SIC}} \\
& + \underbrace{\sum_{m=1}^{M^s} \mathbf{g}_{mnk}^{\mathrm{H}} \sum_{n'=1, n' \neq n}^{N} \mathbf{w}_{mn'}^s s_{n'}^s}_{\text{inter-cluster interference}} \\
& + \underbrace{\sqrt{p_t^p} \mathbf{u}_{nk}^{\mathrm{H}} \sum_{j=1}^{N} \mathbf{w}_j^p \sqrt{\lambda_j^p} s_j^p}_{\text{interference from the primary network}} + n_{nk}^s,
\end{aligned} \tag{9}
$$

where $n_{nk}^s \sim \mathcal{CN}(0,1)$ and $\hat{s}_{ni}^s$ is the estimate of s_{ni}^s. Here, $\hat{s}_{ni}^s$ and s_{ni}^s are jointly Gaussian distributed with a normalized correlation coefficient ρ_{ni} given by $s_{ni}^s = \rho_{ni} \hat{s}_{ni}^s + e_{ni}$, where s_{ni}^s, $\hat{s}_{ni}^s \sim \mathcal{CN}(0,1)$, $e_{ni} \sim \mathcal{CN}\big(0, \frac{\sigma_{e_{ni}}^2}{1+\sigma_{e_{ni}}^2}\big)$ and $\rho_{ni} = \frac{1}{\sqrt{1+\sigma_{e_{ni}}^2}}$. And $\hat{s}_{ni}^s$ and e_{ni} are independent [11].

III. Downlink Rate and User Clustering

We next derive the secondary achievable rate by using the worst-case Gaussian technique [1]. From (9), the achievable rate for the kth SU in the nth cluster can be written as

$$
\mathcal{R}_{nk}^s = \phi \log_2(1 + \gamma_{nk}^s), \tag{10}
$$

where $\phi = (\tau_c - \tau)/\tau_c$ is the pre-log factor and τ_c is the coherence interval. Let γ_{nk}^s be the effective signal-to-interference-plus-noise ratio (SINR) at the kth SU in the nth cluster. To determine it, we consider the first term in (9) to be the desired signal and the remaining terms be an effective noise. Thus, γ_{nk}^s can be derived as

$$
\gamma_{nk}^s = p_d / (p_t^s \sum_{i=1}^{4} P_{I_i} + p_t^p P_{I_5} + 1), \tag{11}
$$

in which

$$
\begin{aligned}
p_d &= p_t^s \lambda_{nk}^s \left|\mathbb{E}\left\{\eta_{nk}^s\right\}\right|^2, \\
P_{I_1} &= \lambda_{nk}^s \mathbb{E}\left\{\left|(\eta_{nk}^s - \mathbb{E}\left\{\eta_{nk}^s\right\})\right|^2\right\}, \\
P_{I_2} &= \sum_{i=1}^{k-1} \lambda_{ni}^s \mathbb{E}\left\{|\eta_{nk}^s|^2\right\}, \\
P_{I_3} &= \sum_{i=k+1}^{K} \lambda_{ni}^s \mathbb{E}\left\{\left|\eta_{nk}^s s_{ni}^s - \mathbb{E}\left\{\eta_{nk}^s\right\} \hat{s}_{ni}^s\right|^2\right\}, \\
P_{I_4} &= \sum_{n'=1, n' \neq n}^{N} \lambda_{n'}^s \mathbb{E}\left\{\left|\eta_{n'k}^s\right|^2\right\}, \\
P_{I_5} &= \sum_{j=1}^{N} \lambda_j^p \mathbb{E}\left\{\left|\mathbf{u}_{nk}^{\mathrm{H}} \mathbf{w}_j^p\right|^2\right\},
\end{aligned} \tag{12}
$$

where $\eta_{nk}^s \triangleq \sum_{m=1}^{M^s} \mathbf{g}_{mnk}^{\mathrm{H}} \mathbf{w}_{mn}^s$.

By evaluating the expectation terms in (12), the effective SINR η_{nk}^s can be derived (Appendix B) as (13), shown at the bottom of this page, where $\zeta_n^p \triangleq 1 + \tau p_p (\sum_{i=1}^{K} \beta_{\mathbf{u}_{ni}} + \beta_{\mathbf{h}_n})$. Then, the achievable sum rate of our system can be expressed as

$$
\mathcal{R}^s = \sum_{n=1}^{N} \sum_{k=1}^{K} \mathcal{R}_{nk}^s, \tag{14}
$$

where $\mathcal{R}_{nk}^s$ is defined in (10) and is calculated to satisfy the NOMA condition [12, eq. (9)].

A. User Clustering

We now propose the user-clustering scheme based upon the Jaccard distance coefficient [10] to find the most dissimilar SUs in the secondary network. We define the centroid point of all the large-scale channel coefficients between the SUs and the secondary APs. The centroid enables us to define the Jaccard coefficients. By using them, we quantify the dissimilarity/similarity of SUs with respect to the centroid.

Let $\boldsymbol{\beta}_k = [\beta_{k1}, \ldots, \beta_{kM^s}]$ contains the large-scale gains of the channels between the kth SU and all the secondary APs,

$$
\gamma_{nk}^s = \frac{L p_t^s \lambda_{nk}^s \left(\sum_{m=1}^{M^s} \sqrt{\theta_{\hat{\mathbf{g}}_{mnk}}}\right)^2}{L p_t^s \left(\sum_{m=1}^{M^s} \sqrt{\theta_{\hat{\mathbf{g}}_{mnk}}}\right)^2 \left(\sum_{i=1}^{k-1} \lambda_{ni}^s + \sum_{i=k+1}^{K} \lambda_{ni}^s (2 - 2\rho_{ni})\right) + p_t^s \left(\sum_{n'=1}^{N} \lambda_{n'}^s\right) \sum_{m=1}^{M^s} \beta_{\mathbf{g}_{mnk}} + p_t^p \left(\left(\sum_{n'=1}^{N} \lambda_{n'}^p\right) \beta_{\mathbf{u}_{nk}} + \lambda_n^p \frac{M^p \tau p_p}{\zeta_n^p} \beta_{\mathbf{u}_{nk}}^2\right) + 1} \tag{13}
$$

Algorithm 1 User-Clustering for CF MIMO-NOMA

1. Compute κ_k $\forall k \in \{1, \ldots, KN\}$ **using** (15).
2. Sort SUs in descending order of similarity:
$\kappa_1 \geq \kappa_2 \geq \cdots \geq \kappa_{KN}$.
3. Group SUs into N **clusters with** K $(K \geq 2)$ **users per cluster:**
1st cluster = $\{\kappa_1, \kappa_{N+1}, \kappa_{2N+1}, \ldots, \kappa_{KN}\}$,
2nd cluster = $\{\kappa_2, \kappa_{N+2}, \kappa_{2N+2}, \ldots, \kappa_{KN-1}\}$,...
Nth cluster = $\{\kappa_N, \kappa_{2N}, \kappa_{3N}, \ldots, \kappa_{KN-(N-1)}\}$.

which specifies the location of the kth SU. Considering the centroid point as $\boldsymbol{\beta}_c = 1/(NK)\sum_{k=1}^{NK}\boldsymbol{\beta}_k$, the dissimilarity metric based on Jaccard coefficient (a.k.a. Tanimoto) can be written as [10]

$$\kappa_k = \frac{\boldsymbol{\beta}_c \boldsymbol{\beta}_k^{\mathrm{T}}}{\|\boldsymbol{\beta}_c\|^2 + \|\boldsymbol{\beta}_k\|^2 - \boldsymbol{\beta}_c \boldsymbol{\beta}_k^{\mathrm{T}}}, \quad \forall k \tag{15}$$

where κ_k ranges from 0 (perfect dissimilarity) to 1 (perfect similarity). We note that Jaccard coefficient captures both the angle difference between vectors (cosine similarity) and the difference in their lengths (Euclidean distance).

Then, the SUs are sorted in a descending order of similarity with the centroid $\boldsymbol{\beta}_c$ i.e., $\kappa_1 \geq \kappa_2 \geq \cdots \geq \kappa_{KN}$ and the SUs which are located far from each other are put in a cluster. **Algorithm 1** summarizes the proposed user-clustering.

IV. Simulation Results

Herein, we provide simulation results to evaluate the performance of our secondary system. For comparisons, we also consider an OMA. The minimum pilot sequence lengths for NOMA and OMA are $\tau_{\text{NOMA}} = N$ and $\tau_{\text{OMA}} = KN$, respectively. The pre-log factors are $\phi_{\text{NOMA}} = (\tau_c - N)/\tau_c$ and $\phi_{\text{OMA}} = (\tau_c - KN)/\tau_c$, respectively.

We consider an area of size 2×2 km^2, and the primary BS is located at the center and the PUs are distributed uniformly at random around it with the minimum and maximum distances of 30 m and 300 m, respectively.[1] The secondary APs and SUs are uniformly distributed in the given area. However, to satisfy the interference constraints, a protected zone is implemented around the BS. This is done as follows. The APs (SUs) closer than 400 m to the BS are turned off by the CPU (not served by the active secondary APs). The served SUs are clustered as per Algorithm 1 (clustering I), and we assume two SUs per cluster ($K = 2$). Uniform power allocation for the primary network ($\lambda_n^p = 1/N$) and secondary APs to each cluster ($\lambda_n^s = 1/N$) is considered. The large-scale coefficients $\{\beta_{\mathbf{a}}\}$ are an uncorrelated shadow fading process with standard deviation $\sigma_{\text{sh}} = 8$ dB [1]. In all simulations, we assume that $M^p = M^s = 100$, $p_t^p = p_{t,\max}^s = 200$ mW, $p_p = 100$ mW, $\lambda_{n1}^s = 0.3\lambda_n^s$, $\lambda_{n2}^s = 0.7\lambda_n^s$ $\forall n$ (5). P-SIC and I-SIC stand for perfect SIC ($\rho_{nk} = 1$) and imperfect SIC ($\rho_{nk} = 0.1$).

Fig. 2 depicts the achievable sum rate (14) of the secondary network with NOMA or OMA as a function of the number of SUs where $\tau_c = 196$. As a benchmark, we also plot the achievable rate of the secondary network when a set of proximate users are grouped as a cluster [4] (legend: clustering II). We see that our proposal effectively exploits the channel gain differences and significantly improves the sum rate. We also observe that, OMA can serve $K_{\max}^{\text{OMA}} = 196$ SUs simultaneously; but NOMA doubles this number of SUs. The reason is that OMA requires different mutually orthogonal pilots while NOMA uses just one per cluster. For a large number of SUs, NOMA outperforms OMA; however, for a small number of SUs, OMA outperforms NOMA due to the effects of intra-cluster pilot contamination and imperfect SIC. We also observe that, imperfect SIC degrades NOMA; for instance, for clustering I, a 1.78 bps/Hz hit occurs for 140 simultaneous SUs when $I_p = 1.5$. Here, we consider two different values $I_p = 1.5, 0.5$ mW (equal I_p for all the PUs). As expected, low values of I_p offer better protection against secondary interference at the PUs, albeit at the expense of the secondary sum rate.

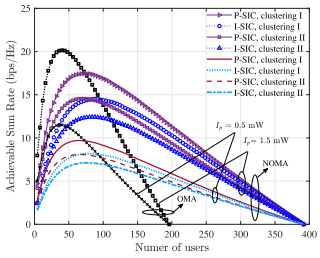

Fig. 2. The achievable sum rate versus the number of users ($L = 5$).

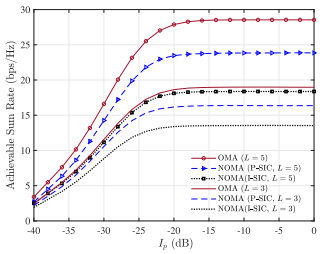

Fig. 3. The achievable sum rate versus I_p.

Fig. 3 shows the achievable sum rate of NOMA and OMA for the secondary network for different values of I_p when $N = 25$. With low I_p, the achieved sum rate increases, while it saturates to a constant in the regime of high values of I_p. This is because strict protection against the secondary interference is maintained and the secondary APs are allowed to transmit with low powers (6), which reduce the sum rates. However, as I_p grows, the secondary APs are allowed to transmit with higher powers and the SUs could achieve higher rates. Once the constraint in (12) are met, secondary APs can transmit with the maximum available power and the sum rate saturates and becomes independent of I_p. The impact of more AP antennas is also investigated in Fig. 3. We observe that the use of more antennas increases the sum rate, thanks to the array gain. However, this also increases the interference due to pilot contamination and imperfect SIC in NOMA (see the denominator of (13)).

V. Conclusion

In this letter, we analyzed the achievable rate of a CF massive MIMO-NOMA (secondary) and massive MIMO (primary) system. We suggested SU clustering based on Jaccord coefficients and derived the closed-form sum rate expression of the secondary network considering joint effects of intra-cluster

[1]The maximum distance is specified in order to guarantee the quality of service of the PUs given the BS peak power constraint and the pathloss.

pilot contamination, inter-cluster interference, imperfect SIC and statistical downlink CSI at SUs. We found that NOMA based cognitive CF massive MIMO can support significantly more SUs compared to the OMA-hybrid.

Finally, we would like to note that, compared to OMA-hybrid, NOMA-hybrid introduces additional hardware complexity due to the SIC processing and error propagation. Essentially, NOMA introduces a tradeoff between performance (sum rate) and complexity. Analyzing this tradeoff is a worthwhile future research direction.

APPENDIX A: DERIVATION OF $P(z_n)$

Regarding (7), we have

$$\begin{aligned} P(z_n) &= \sum_{\substack{n'=1\\ n'\neq n}}^{N} \lambda_{n'}^s \sum_{m=1}^{M^s} \mathbb{E}\left\{\left|\mathbf{v}_{mn}^{\mathrm{H}}\mathbf{w}_{mn'}^s\right|^2\right\} \\ &\quad + \lambda_n^s \sum_{m=1}^{M^s} \mathbb{E}\left\{\left|\mathbf{v}_{mn}^{\mathrm{H}}\mathbf{w}_{mn}^s\right|^2\right\} \\ &\quad + \lambda_n^s \sum_{m=1}^{M^s} \sum_{\substack{m'=1\\ m'\neq m}}^{M^s} \mathbb{E}\left\{\left(\mathbf{v}_{mn}^{\mathrm{H}}\mathbf{w}_{mn}^s\right)\left(\mathbf{v}_{m'n}^{\mathrm{H}}\mathbf{w}_{m'n}^s\right)^{\mathrm{H}}\right\}, \quad \text{where} \end{aligned} \tag{16}$$

$$\begin{aligned} &\mathbb{E}\left\{\left|\mathbf{v}_{mn}^{\mathrm{H}}\mathbf{w}_{mn'}^s\right|^2\right\} \\ &= \frac{1}{L\zeta_{mn'}^s}\left(L\beta_{\mathbf{v}_{mn}}\left(1+\tau p_p\sum_{k=1}^{K}\beta_{\mathbf{g}_{mn'k}}\right)+\tau p_p L\beta_{\mathbf{v}_{mn}}\beta_{\mathbf{v}_{mn'}}\right) \\ &= \frac{1}{\zeta_{mn'}^s}\left(\beta_{\mathbf{v}_{mn}}\zeta_{mn'}^s\right) = \beta_{\mathbf{v}_{mn}}. \end{aligned} \tag{17}$$

We also find

$$\begin{aligned} &\mathbb{E}\left\{\left|\mathbf{v}_{mn}^{\mathrm{H}}\mathbf{w}_{mn}^s\right|^2\right\} \\ &= \frac{1}{L\zeta_{mn}^s}\left(L(L+1)\tau p_p\beta_{\mathbf{v}_{mn}}^2+L\beta_{\mathbf{v}_{mn}}\left(1+\tau p_p\sum_{k=1}^{K}\beta_{\mathbf{g}_{mnk}}\right)\right) \\ &= \left(\beta_{\mathbf{v}_{mn}}+\frac{L}{\zeta_{mn}^s}\tau p_p\beta_{\mathbf{v}_{mn}}^2\right). \end{aligned} \tag{18}$$

Moreover

$$\mathbb{E}\left\{\left(\mathbf{v}_{mn}^{\mathrm{H}}\mathbf{w}_{mn}^s\right)\left(\mathbf{v}_{m'n}^{\mathrm{H}}\mathbf{w}_{m'n}^s\right)^{\mathrm{H}}\right\} = \frac{L\tau p_p}{\sqrt{\zeta_{mn}^s\zeta_{m'n}^s}}\beta_{\mathbf{v}_{mn}}\beta_{\mathbf{v}_{m'n}}, \tag{19}$$

in which $\mathbb{E}\left\{\left(\mathbf{v}_{mn}^{\mathrm{H}}\mathbf{w}_{mn}^s\right)\right\} = \frac{1}{\sqrt{\zeta_{mn}^s}}\sqrt{L\tau p_p}\beta_{\mathbf{v}_{mn}}$. Finally, by substituting (17), (18) and (19) in (16), $P(z_n)$ is given as (8).

APPENDIX B: DERIVATION OF γ_{nk}^s (11)

To find γ_{nk}^s, the following terms need to be derived:

$$\mathbb{E}\left\{\eta_{nk}^s\right\} = \sum_{m=1}^{M^s}\sqrt{L\theta_{\hat{\mathbf{g}}_{mnk}}}, \quad \mathbb{E}\left\{\left|\eta_{n'k}^s\right|^2\right\} = \sum_{m=1}^{M^s}\beta_{\mathbf{g}_{mnk}}, \tag{20}$$

$$\mathbb{E}\left\{\left|\left(\eta_{nk}^s - \mathbb{E}\left\{\eta_{nk}^s\right\}\right)\right|^2\right\} = \sum_{m=1}^{M^s}\beta_{\mathbf{g}_{mnk}}, \tag{21}$$

$$\mathbb{E}\left\{\left|\eta_{nk}\right|^2\right\} = \sum_{m=1}^{M^s}\beta_{\mathbf{g}_{mnk}} + \left(\sum_{m=1}^{M^s}\sqrt{L\theta_{\hat{\mathbf{g}}_{mnk}}}\right)^2, \tag{22}$$

$$\begin{aligned} &\mathbb{E}\left\{\left|\eta_{nk}s_{ni}^s - \mathbb{E}\left\{\eta_{nk}\right\}\hat{s}_{ni}^s\right|^2\right\} = \mathbb{E}\left\{\left|\eta_{nk}\right|^2\right\} \\ &\quad + (1-2\rho_{ni})\mathbb{E}^2\{\eta_{nk}\}. \end{aligned} \tag{23}$$

The proofs of (20) - (23) are omitted due to space limitations, but follows the same principles as [4]. Besides

$$\begin{aligned} &\sum_{j=1}^{N}\lambda_j^p\mathbb{E}\left\{\left|\mathbf{u}_{nk}^{\mathrm{H}}\mathbf{w}_j^p\right|^2\right\} \\ &\quad = \lambda_n^p\mathbb{E}\left\{\left|\mathbf{u}_{nk}^{\mathrm{H}}\mathbf{w}_n^p\right|^2\right\} + \sum_{j=1,j\neq n}^{N}\lambda_j^p\mathbb{E}\left\{\left|\mathbf{u}_{nk}\mathbf{w}_j^p\right|^2\right\}. \end{aligned} \tag{24}$$

where, $\mathbb{E}\left\{\left|\mathbf{u}_{nk}^{\mathrm{H}}\mathbf{w}_j^p\right|^2\right\} = \beta_{\mathbf{u}_{nk}}$ and

$$\begin{aligned} &\mathbb{E}\left\{\left|\mathbf{u}_{nk}^{\mathrm{H}}\mathbf{w}_n^p\right|^2\right\} \\ &= \frac{1}{M^p\zeta_n^p}\left(M^p(M^p+1)\tau p_p\beta_{\mathbf{u}_{nk}}^2\right) \\ &\quad + \frac{1}{M^p\zeta_n^p}\left(M^p\beta_{\mathbf{u}_{nk}}\left(1+\tau p_p\left(\beta_{\mathbf{h}_n}+\sum_{j=1,j\neq k}^{K}\beta_{\mathbf{u}_{nj}}\right)\right)\right) \\ &= \beta_{\mathbf{u}_{nk}} + \frac{M^p}{\zeta_n^p}\tau p_p\beta_{\mathbf{u}_{nk}}^2. \end{aligned} \tag{25}$$

REFERENCES

[1] H. Q. Ngo *et al.*, "Cell-free massive MIMO versus small cells," *IEEE Trans. Wireless Commun.*, vol. 16, no. 3, pp. 1834–1850, Mar. 2017.

[2] S. M. R. Islam *et al.*, "Power-domain non-orthogonal multiple access (NOMA) in 5G systems: Potentials and challenges," *IEEE Commun. Surveys Tutr.*, vol. 19, no. 2, pp. 721–742, 2nd Quart., 2017.

[3] M. Vaezi *et al.*, "Interplay between NOMA and other emerging technologies: A survey," *IEEE Trans. Cogn. Commun. Netw.*, vol. 5, no. 4, pp. 900–919, Dec. 2019.

[4] Y. Li and G. A. Aruma Baduge, "NOMA-aided cell-free massive MIMO systems," *IEEE Wireless Commun. Lett.*, vol. 7, no. 6, pp. 950–953, Dec. 2018.

[5] H. Al-Hraishawi *et al.*, "Multi-cell massive MIMO uplink with underlay spectrum sharing," *IEEE Trans. Cogn. Commun. Netw.*, vol. 5, no. 1, pp. 119–137, Mar. 2019.

[6] S. Kusaladharma and C. Tellambura, "Secondary user interference characterization for spatially random underlay networks with massive MIMO and power control," *IEEE Trans. Veh. Technol.*, vol. 66, no. 9, pp. 7897–7912, Sep. 2017.

[7] D. L. Galappaththige and G. Amarasuriya, "Cell-free massive MIMO with underlay spectrum-sharing," in *Proc. IEEE Int. Conf. Commun. (ICC)*, May 2019, pp. 1–7.

[8] E. Björnson *et al.*, "Massive MIMO: Ten myths and one critical question," *IEEE Commun. Mag.*, vol. 54, no. 2, pp. 114–123, Feb. 2016.

[9] M. S. Ali *et al.*, "Non-orthogonal multiple access (NOMA) for downlink multiuser MIMO systems: User clustering, beamforming, and power allocation," *IEEE Access*, vol. 5, pp. 565–577, 2017.

[10] S. H. Cha, "Comprehensive survey on distance/similarity measures between probability density functions," *Int. J. Math. Models Methods Appl. Sci.*, vol. 1, no. 4, pp. 300–307, Nov. 2007.

[11] S. M. Kay, *Fundamentals Statistical Signal Processing: Estimation Theory*. Englewood Cliffs, NJ, USA: Prentice-Hall, 1993.

[12] M. F. Hanif *et al.*, "A minorization-maximization method for optimizing sum rate in the downlink of non-orthogonal multiple access systems," *IEEE Trans. Signal Process.*, vol. 64, no. 1, pp. 76–88, Jan. 2016.