

Access Point Cooperation Strategies for Coded Random Access in Cell-Free Massive MIMO

Enrico Testi¹, *Member, IEEE*, Velio Tralli², *Senior Member, IEEE*, and Enrico Paolini¹, *Senior Member, IEEE*

Abstract—In this article, grant-free uplink communication from a large number of machine-type devices in cell-free massive MIMO networks is explored. A novel approach that leverages coded random access (CRA), on the device side, with combining of signals received at properly selected access points (APs) and cooperative successive interference cancellation (SIC), on the network side, is presented. Initially, an analytical framework based on stochastic geometry is developed to investigate performance of AP cooperation through signal combining under diverse AP cluster compositions. The potential gain from AP signal combining is then assessed by evaluating a genie-aided scheme, guiding the network in cluster selection for each active device. Subsequently, two practical AP selection algorithms that operate in grant-free conditions (i.e., do not require prior information regarding the active users) are proposed. Numerical results show how AP cooperation through signal combining and distributed interference cancellation can bring tangible benefits even without prior information about active users, under different signal-to-noise ratio regimes, closing in some cases the gap to the genie-aided approach. Additionally, the results prove that AP cooperation can be used to reduce the devices' energy consumption and the number of APs that have to be deployed by the service providers to achieve specific performance levels.

Index Terms—AP cooperation, cell-free massive MIMO (CF-mMIMO), CRA, Internet of Things (IoT), massive multiple access (MMA).

I. INTRODUCTION

THE EVER-GROWING integration of the Internet of Things (IoT) in several application domains has prompted an explosion of the number of connected smart objects, leading to the rise of massive machine-type communications (mMTC) [1], [2]. In 6G, mMTC systems will need to meet scalability and device battery lifetime requirements more stringent than 5G ones, with use case-specific reliability and latency levels [3], [4]. In this context, enhanced medium

access protocol (MAC) and physical (PHY) layer schemes for next-generation massive multiple access (MMA), as well as innovative network architectures, will be paramount to address the challenges posed by the growing network densities and the heterogeneous service requirements.

Concerning innovative architectures, cell-free massive MIMO (CF-mMIMO) networks have recently emerged as a promising alternative to traditional, cell-based ones [5], [6], [7], [8], [9]. This architectural concept diverges from a classical cell-based structure in that a massive number of access points (APs) is distributed across the network area. The AP operations are coordinated by one or more central processing units (CPUs) with powerful computation capabilities. This way, the system is shifted from a “base station (BS)-centric” model to a “user-centric” one, where each user accesses the network through the cluster of APs that are at closest distance from it, to constantly maximize its quality of service [10]. The decentralized nature of CF-mMIMO networks, where each AP is equipped with a relatively small number of antennas, leverages the benefits of massive MIMO, achieving high multiplexing and array gains through distributed signal processing. The relatively short links between devices and APs facilitate more effective and equitable coverage, thereby enabling connected devices to operate at lower transmit power levels, a regime of great interest in mMTC. Such network architectures, augmented by emerging technologies like reconfigurable intelligent surfaces (RIS), extremely large MIMO (XL-MIMO), and holographic MIMO, are among the most promising candidates for next-generation wireless networks [11], [12], [13].

In terms of innovative grant-free access protocols supporting MMA, coded random access (CRA) ones [14] have recently gained interest due to their capability to achieve higher levels of reliability (with respect to current mMTC) even with a large number of active uncoordinated devices (e.g., [15], [16], [17], [18], and [19]). Based on the idea of combining packet diversity with successive interference cancellation (SIC) across various resources at the receiver, these protocols can provide remarkable scalability gains [14]. In the CRA protocol considered in this work, an active user transmits multiple replicas of its data payload in different time slots; whenever any such replica is decoded in a slot, the interference generated by the decoded packet and by its twins in the corresponding slots is subtracted from the received signal samples. Reprocessing the signal in these slots can potentially result in the decoding of additional messages. Bridging massive access

Manuscript received 15 March 2024; revised 29 May 2024 and 2 July 2024; accepted 5 July 2024. Date of publication 15 July 2024; date of current version 23 August 2024. This work was supported in part by the Consorzio Nazionale Interuniversitario per le Telecomunicazioni (CNIT) Wireless Communications Laboratory (WiLab) and the WiLab-Huawei Joint Innovation Center, and in part by the European Union through the Italian National Recovery and Resilience Plan of NextGenerationEU, partnership on “Telecommunications of the Future” under Grant PE000000001-RESTART. This article was presented in part at the 2024 IEEE International Conference on Communications (ICC 2024) Workshops. (Corresponding author: Enrico Testi.)

Enrico Testi and Enrico Paolini are with the CNIT/WiLab, Department of Electrical, Electronic and Information Engineering, University of Bologna, 47522 Cesena, Italy (e-mail: enrico.testi@unibo.it; e.paolini@unibo.it).

Velio Tralli is with the CNIT/WiLab, Department of Engineering, University of Ferrara, 44121 Ferrara, Italy (e-mail: velio.tralli@unife.it).

Digital Object Identifier 10.1109/IIOT.2024.3428922

with modern codes-on-graphs, these protocols have so far been studied mainly in a centralized context, where all active users contend to deliver their messages to the same BS. To the best of the author's knowledge, the only work considering CRA in cell-free networks is [21].

In the currently available scientific literature, although massive access is a predominant use case in the context of CF-mMIMO systems, only a few facets of it have been investigated. In [22] and [23], activity detection was tackled as a maximum likelihood (ML) estimation problem, while [24] delved into joint activity detection, channel estimation, and data estimation via belief propagation. Distributed and centralized processing for joint activity detection and channel estimation with nonorthogonal random pilot sequences were discussed in [25]. Strategies aimed at minimizing the impact of pilot contamination in CF-mMIMO networks through power control and pilot assignment have been proposed recently [26], [27], [28], [29], [30]. In both [27] and [28], a preliminary phase during which a generic user equipment (UE) is associated with a subset of available APs, is followed by a practical strategy to assign orthogonal pilots to UEs aiming at reducing the mutual interference and improving the spectral efficiency. Conversely, Rahmani et al. [29] assumed that all APs receive the signals of all UEs and develop a scalable and distributed pilot assignment mechanism via multiagent reinforcement learning. In [26] and [30], power control is formulated as an optimization problem to reduce the effect of pilot contamination; the objective of [30] is to maximize the minimum signal-to-interference-plus-noise ratio (SINR), while that of [26] is to enforce a desired minimum SINR across all the UEs. In addition, [31], [32] focused on the downlink. In [21], a CRA scheme tailored to CF-mMIMO networks was described. Here, the APs process the received signals individually while the CPU orchestrates a distributed SIC procedure with very low overhead on the fronthaul links.

In this article, following the preliminary work in [20], differently from [21], we investigate the possibility of enhancing the performance of massive grant-free access in CF-mMIMO networks (with reference to a CRA transmission scheme) through a judicious selection of subsets of APs, whose signals are combined to detect and decode the received messages. Owing to the assumed fully grant-free nature of the access protocol, the AP selection must be performed at the CPU level based solely on the signal samples received at the APs; this leads to a compelling challenge since no foreknowledge can be assumed about the number of active users, their spatial positions, and the pilot each of them employs.

The main contributions of this work are the following.

- 1) We introduce and systematically analyze different levels of cooperation among APs in grant-free CF-mMIMO networks, exploring how they can influence network performance. This includes the novel concept of cooperative detection and decoding of user messages through dynamically formed clusters of APs.
- 2) We derive an analytical framework based on stochastic geometry to study the performance of AP cooperation in packet detection and decoding under various AP cluster compositions.

- 3) We analyze the potentially attainable gain of cooperative detection and decoding using a genie-aided approach, where AP clusters are optimally selected based on the full knowledge of the system. This provides a benchmark to evaluate the effectiveness of practical cooperative schemes, thereby motivating the development of efficient cooperation mechanisms.
- 4) We propose two novel schemes for "uninformed" cluster formation, i.e., cluster formation where the system operates without prior information about active users, apart from a set of available pilots that they can choose individually, in a grant-free manner.
- 5) Finally, extensive simulations are carried out to assess the performance of the proposed AP cooperation schemes. These simulations compare the proposed methods against noncooperative and basic cooperative schemes. Additional considerations regarding the possibility of exploiting AP cooperation to reduce the devices' energy consumption and the number of APs deployed in the service area are presented.

The system model and the CRA-based processing schemes with no AP signal combining are presented in Section II. The analytical framework for the performance analysis of AP signal combining is presented in Section III. The genie-aided and practical AP cooperation strategies are described in Section IV. Numerical results are presented in Section V and conclusions are drawn in Section VI.

Notation: Boldface lower case and upper case letters denote row vectors and matrices, respectively. The operations $(\cdot)^T$, and $(\cdot)^H$ indicate transpose and conjugate transpose, respectively. The expectation, variance, and Euclidean norm operators are defined as $\mathbb{E}[\cdot]$, $\mathbb{V}[\cdot]$, and $\|\cdot\|$, respectively. The normal and circularly symmetric complex normal distributions with mean 0 and variance σ^2 are denoted by $\mathcal{N}(0, \sigma^2)$ and $\mathcal{CN}(0, \sigma^2)$, respectively. The cosecant function is denoted by $\text{csc}(\cdot)$. We also let $[n] = \{1, \dots, n\}$.

II. SYSTEM MODEL AND BACKGROUND

We consider a CF-mMIMO system with M APs, each with N antennas, deployed in an area. The APs are connected to a CPU via fronthaul links, as depicted in Fig. 1, and jointly serve a massive number of machine-type devices (not all simultaneously active) under CPU coordination on the same time-frequency resources. In the figure, the APs are arranged on a grid although this is not required by the developed algorithms. Uplink is synchronous, i.e., there is a common time reference between devices, APs, and CPU. The time is segmented into frames, each divided into N_s slots. Time synchronization is achieved through a beacon signal transmitted by the APs at the start of each frame. Such synchronization signals are standard in cellular networks, such as the primary and secondary synchronization signals in 5G [5]. Let us also remark that, while perfect time synchronization may be challenging in large networks due to propagation delays, if each AP is synchronized with its neighboring APs and users synchronize with the APs in their immediate vicinity then our proposed system model remains valid [5]. After waking

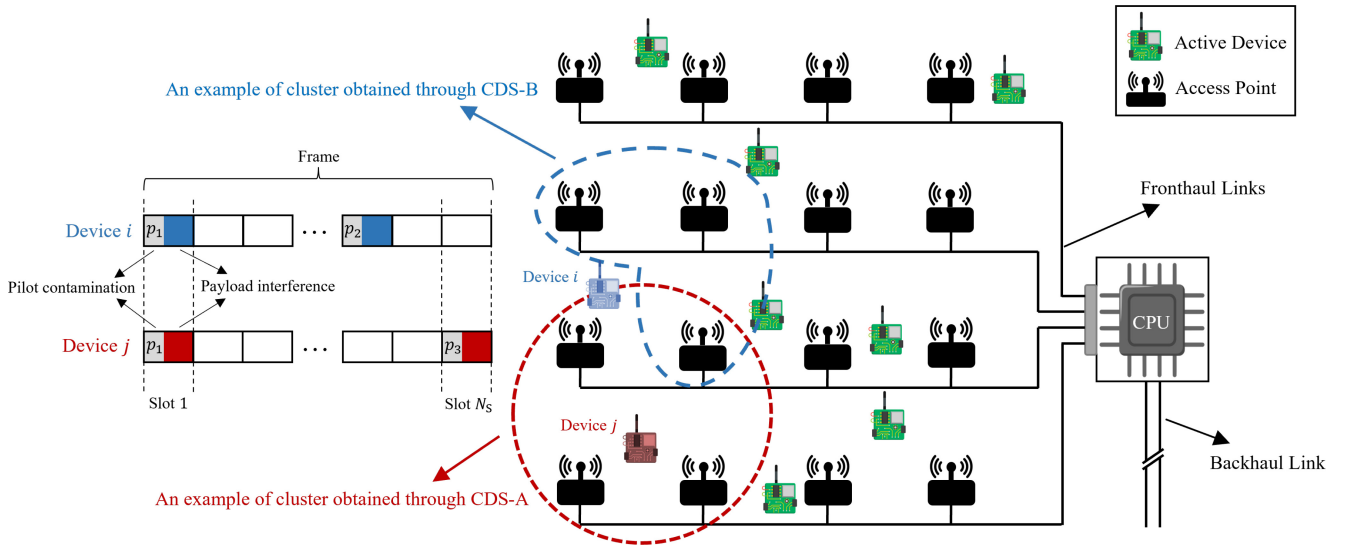


Fig. 1. Illustration of a CRA CF-mMIMO scenario where two clusters of APs obtained via CDS-A and CDS-B strategies perform cooperative detection and decoding of the packets transmitted by the j th and the i th active devices, respectively.

up, an active device waits for the beacon and then competes for transmission of one k -bit message $\mathbf{W} \in [2^k]$ within the subsequent frame.

Let us consider a generic frame and assume to have K devices active over the frame which are randomly distributed in the network area according to a uniform spatial distribution. All devices encode their messages with the same channel encoder and map the coded bits onto the same complex constellation, obtaining a data payload $\mathbf{x} \in \mathbb{C}^{1 \times N_D}$ of N_D complex symbols. An active device selects R slots in $[N_S]$ and transmits a packet $[\mathbf{p}, \mathbf{x}]$ in each such slot, where $\mathbf{p} \in \mathbb{C}^{1 \times N_P}$ is a pilot sequence drawn from a set $\mathcal{P}$ of $P \ll K$ mutually orthogonal pilots. Note that, while the same payload $\mathbf{x}$ is transmitted in all the R picked slots, the pilot $\mathbf{p}$ may change in each such slot. Both the R slot indexes and the R pilots are generated as functions of the random message $\mathbf{W}$ (e.g., through the algorithm described in [33]). Although the pilot may change, we refer to the transmission of a user in each of its R chosen slots as a “replica.” Each replica arriving at an AP is assumed perfectly aligned with one of the frame slots.

All devices employ the same transmit power ρ . Consider a generic slot and let $\mathcal{U}$ be the subset of active users transmitting a replica in it. Each user in $\mathcal{U}$ first transmits the pilot corresponding to its message. The corresponding vector of signal samples received by antenna n of AP m , $\mathbf{y}_{mn,P} \in \mathbb{C}^{1 \times N_P}$, is

$$\mathbf{y}_{mn,P} = \sqrt{\rho} \sum_{i \in \mathcal{U}} h_{mni} \mathbf{p}(\mathbf{W}_i) + \mathbf{w}_{mn} \quad (1)$$

where $\mathbf{p}(\mathbf{W}_i) \in \mathbb{C}^{1 \times N_P}$ is the pilot function of the message $\mathbf{W}_i$ transmitted by user i , h_{mni} is the channel gain between antenna n of AP m and device i , and $\mathbf{w}_{mn} \sim \mathcal{CN}(\mathbf{0}_{N_P}, \sigma_N^2 \mathbf{I}_{N_P \times N_P})$ is a vector of noise samples. Pilot transmissions are followed by payload transmissions. The corresponding vector of received signal samples at antenna n of AP m , $\mathbf{y}_{mn,D} \in \mathbb{C}^{1 \times N_D}$, is

$$\mathbf{y}_{mn,D} = \sqrt{\rho} \sum_{i \in \mathcal{U}} h_{mni} \mathbf{x}(\mathbf{W}_i) + \mathbf{v}_{mn} \quad (2)$$

where $\mathbf{x}(\mathbf{W}_i)$ is the data payload of message $\mathbf{W}_i$ transmitted by user i , and $\mathbf{v}_{mn} \sim \mathcal{CN}(\mathbf{0}_{N_D}, \sigma_N^2 \mathbf{I}_{N_D \times N_D})$ is a vector of noise samples. The coefficients h_{mni} , for each user i , AP m and antenna n , are the complex channel gains, which are constant over an entire slot and are defined according to a Rayleigh block fading channel model.

The channel gains are modeled as $h_{mni} = g_{mni} \sqrt{\beta_{mi}}$, where $g_{mni} \sim \mathcal{CN}(0, 1)$ and β_{mi} are the small-scale and large-scale fading coefficients, respectively. We have

$$\beta_{mi} = \frac{10^{\sigma_S u_{mi}/10}}{\text{PL}_{mi}} \quad (3)$$

where PL_{mi} is the path-loss between user i and AP m , σ_S is the shadowing intensity, and $u_{mi} \sim \mathcal{N}(0, 1)$. For each (m, n, i) , g_{mni} remains constant through a slot and varies independently from slot to slot, while β_{mi} stays constant over the whole frame. We let g_{mni} and g_{pqj} be independent for any $(m, n, i) \neq (p, q, j)$. Similarly, $\beta_{m,i}$ and $\beta_{p,j}$ are independent for any $(m, i) \neq (p, j)$. In this work we will utilize two distinct path-loss models: 1) a general path-loss model for analytically deriving (in Section III-A) a performance metric useful to evaluate the interplay between pilot contamination and diversity gain when forming clusters of cooperating APs and 2) a specific path-loss model, closely aligned with the 3GPP Urban Microcell Model, for evaluating (in Section V-B) the performance of the proposed processing schemes.

A. Noncooperative CRA

We first delineate a processing scheme that does not foresee any kind of cooperation among the APs, which we refer to as noncooperative (NC). In this scheme, which is obtained as a simplified noncooperative version of the scheme proposed in [21], each AP independently attempts to decode the transmitted packets using only its own received signal, without exchanging information with other APs or the CPU. In NC scheme, the SIC process only mitigates interference from

packets detected locally. Additionally, compared to the scheme in [21], we adapt the processing method to accommodate multiple available pilots and multiple antennas at the APs.

The NC processing is logically split into two phases, namely, an initial slot-by-slot signal detection and decoding followed by a SIC process across the slots, both individually executed by each AP. The two phases are described as follows.

1) *Slot-by-Slot Decoding*: In the generic slot $s \in [N_s]$ (in the following we omit index s to keep the symbol notation light), each AP first processes the received pilot signal samples. The pilot signal in (1) received at antenna n of AP m is projected along pilot j , $j \in [P]$, and ML channel estimation is obtained as

$$\hat{h}_{mn}(j) = \frac{y_{mn,P} p_j^H}{\sqrt{\rho} \|p_j\|^2}. \quad (4)$$

Then, maximal ratio combining (MRC) is performed on the payload samples using the estimate in (4), as

$$z_m(j) = \frac{\hat{h}_m^H(j) Y_{m,D}}{\|\hat{h}_m(j)\|^2} \quad (5)$$

where $\hat{h}_m(j) = [\hat{h}_{m1}(j), \dots, \hat{h}_{mN}(j)]^T$ is the vector of channel gains estimated over pilot j , and $Y_{m,D} = [y_{m1,D}^T, \dots, y_{mN,D}^T]^T$ is the matrix of aggregated received payload symbols at all AP antennas. If the message, sent by a device in the processed slot using pilot j , is retrieved upon successful demapping and decoding performed on $z_m(j)$, the message is pushed to a local list $\mathcal{L}_m$ located at the AP m . This operation is repeated by each AP in all slots and for all pilots.

2) *SIC*: In NC processing, the m th AP processes the messages in its list $\mathcal{L}_m$ in order. Let us denote by W the generic message in the list, and by $[N_s]_W \subseteq [N_s]$ the subset of slots where the replicas of the message W have been transmitted. The m th AP picks the first message W in the list and performs the subtraction of the interference generated by the transmitted replicas of the decoded message, followed by a new decoding attempt, in all slots in which the replicas were transmitted. Since the set $[N_s]_W$, the data payload and the pilot are all functions of the message W only, the AP can retrieve all of this information from it. In particular, for each slot in $[N_s]_W$, AP m first re-estimates the channel using the known payload, $x(W)$, as

$$\tilde{h}_{mn}^{(W)} = \frac{y_{mn,D} x(W)^H}{\|x(W)\|^2} \quad (6)$$

and then subtracts $\tilde{h}_{mn}^{(W)} [p(W), x(W)]$ from the received pilot and payload samples, $[y_{mn,P}, y_{mn,D}]$. After interference cancellation, the AP attempts decoding of new messages by performing (4) and (5) in each of the processed slots. When a new message is decoded, it is pushed into the bottom of $\mathcal{L}_m$. The pseudocode for NC scheme is detailed in Algorithm 1, where $Y_{m,P} = [y_{m1,P}^T, \dots, y_{mN,P}^T]^T$ is the aggregate of the pilot signals received by all the antennas of AP m .

B. CRA With Cooperative SIC

In this section we review the processing scheme in [21], for the case of multiple available pilots and multiple antennas

Algorithm 1: Pseudo-code for NC Scheme

Phase 1 - Slot-by-Slot Decoding (SD)

```

1 for  $m \leftarrow 1$  to  $M$  do
2   forall  $s \in [N_s]$  do
3     Process slot  $s$ 
4     for  $j \leftarrow 1$  to  $P$  do
5        $\hat{h}_m(j) \leftarrow Y_{m,P} p_j^H / (\sqrt{\rho} \|p_j\|^2)$ 
6        $z_m(j) \leftarrow \hat{h}_m^H(j) Y_{m,D} / \|\hat{h}_m(j)\|^2$ 
7       Attempt decoding  $z_m(j)$ 
8       if decoding successful then
9          $\mathcal{L}_m \leftarrow \{\mathcal{L}_m, W\}$ 
10      end
11    end
12  end
13 end
```

SD($[N_s], m$)

Phase 2 - SIC

```

14 for  $m \leftarrow 1$  to  $M$  do
15   forall  $W \in \mathcal{L}_m$  do
16     forall  $s \in [N_s]_l$  do
17       Process slot  $s$ 
18       for  $n \leftarrow 1$  to  $N$  do
19          $\tilde{h}_{mn}^{(W)} \leftarrow y_{mn,D} x(W)^H / \|x(W)\|^2$ 
20          $[y_{mn,P}, y_{mn,D}] \leftarrow$ 
21            $[y_{mn,P}, y_{mn,D}] - \tilde{h}_{mn}^{(W)} [p(W), x(W)]$ 
22       end
23     end
24     Remove  $W$  from  $\mathcal{L}_m$ 
25     Execute SD( $[N_s]_W, m$ )
26   end
27 end
```

IC(s, m)

at the APs, which we refer to as CRA with cooperative SIC (CS). Differently from the NC scheme, it includes a simple form of cooperation among APs which share all the decoded packets with the help of the CPU, thus enabling a distributed global SIC process. While CS is fundamentally based on noncooperative signal detection at the APs, it incorporates a mild mechanism of cooperation between the APs during SIC, that allows us to demonstrate the incremental benefits of cooperation. In Sections III and IV, we will introduce a superior level of cooperation, which we refer to as cooperative detection and decoding, and propose a set of practical cooperation schemes.

The CS scheme is still logically split into two phases, namely, an initial slot-by-slot processing executed individually by each AP followed by a distributed SIC process. The main difference between NC and CS schemes lies on the second phase: while in NC each AP performs SIC of only the messages it has decoded locally, in CS a distributed cooperative SIC takes place across all the APs. The CS processing is outlined as follows.

1) *Slot-by-Slot Decoding*: Each AP processes the received signals as in the first phase of the NC scheme. At the end of the procedure, each AP forwards its local list of decoded messages to the CPU via fronthaul links; in this way, all messages successfully decoded in the frame are delivered to the CPU that stores them in a global list $\mathcal{L}$.

Algorithm 2: Pseudo-code for CS Scheme**Phase 1 - Slot-by-Slot Decoding (SD)**

```

1 for  $m \leftarrow 1$  to  $M$  do
2   | Execute SD( $[N_s], m$ )
3 end

```

Phase 2 - Cooperative SIC

```

4 Merge all the lists  $\mathcal{L}_m$  for  $m = 1, \dots, M$ 
5 forall  $W \in \mathcal{L}$  do
6   for  $m \leftarrow 1$  to  $M$  do
7     forall  $s \in [N_s]_W$  do
8       | Process slot  $s$ 
9       | IC( $s, m$ )
10    end
11  end
12  Remove  $W$  from  $\mathcal{L}$ 
13  Execute SD( $[N_s]_W, m$ ),  $m = 1, \dots, M$ 
14 end

```

} DIC($[N_s]_W$)

2) *Cooperative SIC*: The CPU processes the messages in the list $\mathcal{L}$ in order. At each step, the CPU picks the first message in the list and sends it to all the APs for SIC. Let us denote by W such message. Since the data payload, the pilot and the set $[N_s]_W$ are all functions of the message only, each AP can retrieve all this information from it. Each AP, for each slot in $[N_s]_W$, first re-estimates the channel from the payload, then performs subtraction of the interference generated by the transmitted replicas of the decoded message, as detailed in phase 2 of NC scheme. After interference cancelation, the AP attempts decoding of new messages by performing (4) and (5) in each of the processed slots. When a new message is decoded, it is sent to the CPU via fronthaul link and added at the bottom of $\mathcal{L}$. The pseudocode for CS scheme is detailed in Algorithm 2.

III. COOPERATIVE DETECTION AND DECODING

In this section, we first introduce cooperative detection and decoding of the users' messages through clusters of APs. Subsequently, we investigate the performance of AP signal combining for distinct cluster compositions and under different network load conditions. Note that cooperation is here meant as a combining operation, performed by the CPU on the signal samples of suitably defined AP subsets. Although selecting large sets of APs might appear the best approach at first glance, this is actually not always the case: APs with an unfavorable SINR or receiving contaminated pilots, when added to a set of cooperating APs may in fact jeopardize successful message decoding. This effect is even more exacerbated under high-network load conditions. Let us also remark that the APs cluster construction is carried out at the CPU level, relying solely on the signal samples received at the APs. This poses a significant challenge, as it has to be performed in a fully grant-free scenario where no prior information is assumed regarding the number of active users, their spatial positions, or the specific pilot each user employs.

In order to clarify how cooperative detection and decoding is implemented and which issues arise when building a subset

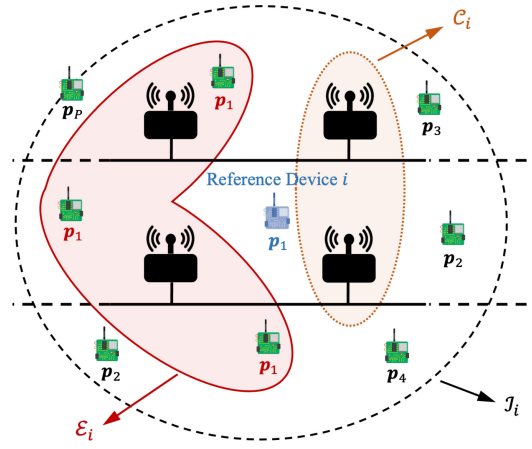


Fig. 2. Graphical illustration of the sets of devices and APs introduced in Section III: 1) reference device i , which is active and is transmitting on the slot s by using pilot p_1 ; 2) the set of APs C_i , that can be potentially used to cooperatively decode the packet replica transmitted by device i in a generic slot s ; 3) the set $\mathcal{I}_i = \mathcal{U} \setminus \{i\}$ of devices transmitting a packet replica in the considered slot, except the i th device; and 4) the set $\mathcal{E}_i \subset \mathcal{I}_i$, i.e., the subset of devices transmitting pilot p_1 .

of cooperating APs, let us first focus on the detection and decoding of a given message W_i of user i , assuming that we know the user is active and is transmitting in slot s using pilot $p(W_i)$. Let us denote by $C_i = \{t_1, \dots, t_{C_i}\}$ the set of APs with cardinality $C_i = |C_i|$ that can be potentially used to cooperatively decode the packet replica transmitted by device i in slot s (for simplicity, we omit slot index). Let us also denote by $\mathcal{I}_i = \mathcal{U} \setminus \{i\}$ the set of devices transmitting a packet replica in the considered slot, except the i th one, and by $\mathcal{E}_i \subset \mathcal{I}_i$ the subset of devices transmitting pilot $p(W_i)$. Fig. 2 provides a graphical illustration of the introduced sets of devices and APs.

Cooperative detection and decoding of message W_i is based on the linear combination of the received payload signals of all the C_i APs forming cluster C_i . If MRC is used to combine the signals, decoding of the user's message is attempted from vector $\mathbf{v}_i \in \mathbb{C}^{1 \times N_D}$ obtained as

$$\mathbf{v}_i = \frac{\hat{\mathbf{q}}_i^H \tilde{\mathbf{Y}}_{i,D}}{\|\hat{\mathbf{q}}_i\|^2} \quad (7)$$

where $\tilde{\mathbf{Y}}_{i,D} = [\mathbf{Y}_{t_1,D}^T, \dots, \mathbf{Y}_{t_{C_i},D}^T]^T$, with $\mathbf{Y}_{m,D} = [\mathbf{y}_{m1,D}^T, \dots, \mathbf{y}_{mN,D}^T]^T$, $m \in C_i$, is the aggregate of the payload signals received by all the antennas of the APs in cluster C_i , and the estimates of all channel gains are stacked in the single vector $\hat{\mathbf{q}}_i = [\hat{\mathbf{h}}_{t_1,i}^T, \dots, \hat{\mathbf{h}}_{t_{C_i},i}^T]^T$, where $\hat{\mathbf{h}}_{mi} = [\hat{h}_{m1i}, \dots, \hat{h}_{mN_i}]$ with $m \in C_i$. Note that the MRC operation in (7), performed by the CPU with the signals of the cooperating APs, is obtained by stacking all their channel coefficients and received signals, even though the APs are not co-located.

The effectiveness of signal combining in improving the signal-to-noise ratio (SNR) and reducing the interference power depends on the quality of the channel gains estimated by the APs, as well as on the composition of cluster C_i , which must be appropriately constructed to include the APs receiving signals with minimal interference and exhibiting high-quality channel estimates. In this regard, let us consider the estimated channel gain at antenna n of AP m related to user i , obtained

through ML estimation over the pilot $\mathbf{p}(\mathbf{W}_i)$, as

$$\begin{aligned}\hat{h}_{mni} &= \frac{\mathbf{y}_{mni, \mathbf{P}} \mathbf{p}^H(\mathbf{W}_i)}{\sqrt{\rho} \|\mathbf{p}(\mathbf{W}_i)\|^2} \\ &= h_{mni} + \sum_{j \in \mathcal{E}_i} h_{mnj} + \frac{\mathbf{w}_{mn} \mathbf{p}^H(\mathbf{W}_i)}{\sqrt{\rho} \|\mathbf{p}(\mathbf{W}_i)\|^2}.\end{aligned}\quad (8)$$

We denote as e_{mni} the channel estimation error due to pilot contamination and noise experienced by the n th antenna of the m th AP and related to user i , i.e.,

$$e_{mni} = \sum_{j \in \mathcal{E}_i} h_{mnj} + \frac{\mathbf{w}_{mn} \mathbf{p}^H(\mathbf{W}_i)}{\sqrt{\rho} \|\mathbf{p}(\mathbf{W}_i)\|^2}.\quad (9)$$

The pilot contamination and the noise terms are two independent zero-mean complex Gaussian random variables, that are constant within each slot. The channel estimation bias term in (9), can hinder the capability of the CPU to decode the message transmitted by device i , by reducing the effectiveness of signal combining. In the following section we will analyze the different components of the combined signal in (7) and we will define a performance metric able to capture the joint effects of pilot contamination, payload interference and noise on the quality of the combined signal employed to decode a message.

A. Performance Analysis

The following analysis is focused on a generic slot of the frame. We first calculate the statistical power of the channel estimation error, which will be exploited subsequently. As the active devices transmitting a packet replica in the designated slot, forming the set $\mathcal{I}_i$, are uniformly distributed across the area, it is reasonable to consider them as forming a Poisson point process (PPP) with intensity λ in $\mathbb{R}^2$. Assuming each device selects a pilot randomly, the count of users transmitting pilot $\mathbf{p}(\mathbf{W}_i)$, forming the set $\mathcal{E}_i$, can still be modeled as a PPP with intensity $\lambda_E = \lambda/P$, as a direct consequence of the independent thinning of Poisson processes [34].

To make the analysis valid also when transmitters and receivers are closer to one another, we adopt a bounded path-loss model having the form

$$\text{PL}_{mi} = \mathcal{L}_0(1 + r_{mi}^\alpha)\quad (10)$$

where r_{mi} is the distance between the m th AP and the i th device, α is the path-loss exponent, and $\mathcal{L}_0$ is the minimum path-loss, so that $2\mathcal{L}_0$ is the path-loss at reference distance $r_0 = 1\text{ m}$ [35], [36]. We also define the scaled shadowing intensity parameter as $\tilde{\sigma}_S = \log(10)\sigma_S/(10\sqrt{2})$.

Let us now denote as $\xi = \xi(\lambda_E)$ the average pilot contamination power for each antenna n of AP m , defined as the expectation of the squared amplitude of the sum of the channel coefficients between antenna n of AP m and the interfering devices in $\mathcal{E}_i$, i.e.,

$$\xi(\lambda_E) = \mathbb{E}\left[\left|\sum_{j \in \mathcal{E}_i} h_{mnj}\right|^2\right].\quad (11)$$

The expectation in (11) is taken over all the sources of randomness. We can derive a closed-form expression for $\xi(\lambda_E)$ as follows.

Theorem 1: The expected squared amplitude of the sum of the channel coefficients between the interferers distributed according to a PPP with density λ_E in $\mathbb{R}^2$ and the n th antenna of the m th AP is independent of m and n and is given by

$$\xi(\lambda_E) = \frac{2\pi^2\lambda_E 10^{\tilde{\sigma}_S^2}}{\mathcal{L}_0\alpha} \csc\left(\frac{2\pi}{\alpha}\right).\quad (12)$$

Proof: Let us start by taking the expectation with respect to the small-scale fading. Recalling that $h_{mni} \sim \mathcal{CN}(0, \beta_{mi})$, we have

$$\mathbb{E}_h\left[\left|\sum_{j \in \mathcal{E}_i} h_{mnj}\right|^2\right] = \sum_{j \in \mathcal{E}_i} \beta_{mj}.\quad (13)$$

We thus recognize that $z = \sum_{j \in \mathcal{E}_i} \beta_{mj}$ is a sum of marks, β_{mj} , of a marked PPP with density λ_E [34]. Let us now apply Campbell's theorem to find the moment generating function (m.g.f.) of z as [37, p. 108]

$$\begin{aligned}m_z(t) &= \mathbb{E}[e^{zt}] \\ &= \exp\left(2\pi\lambda_E \int_0^\infty \int (e^{t\beta} - 1) f_\beta(\beta; r) d\beta dr\right).\end{aligned}\quad (14)$$

Recalling that $\mathbb{E}[z] = (dm_z(t)/dt)|_{t=0}$, we have

$$\mathbb{E}[z] = 2\pi\lambda_E \int_0^\infty \mathbb{E}_u[\beta; r] r dr\quad (15)$$

where $u \sim \mathcal{N}(0, 1)$ and the mark β is a function of u . Therefore

$$\begin{aligned}\mathbb{E}_u[\beta; r] &= \int_{-\infty}^\infty \beta f_u(u) du \\ &= \int_{-\infty}^\infty \frac{10^{\sigma_S u/10}}{\mathcal{L}_0(1+r^\alpha)} \frac{1}{\sqrt{2\pi}} e^{-u^2/2} du \\ &= \frac{10^{\log(10)^2\sigma_S^2/(10\sqrt{2})^2}}{\mathcal{L}_0(1+r^\alpha)} \\ &= \frac{10^{\tilde{\sigma}_S^2}}{\mathcal{L}_0(1+r^\alpha)}.\end{aligned}\quad (16)$$

Incorporating (16) into (15) yields

$$\begin{aligned}\xi(\lambda_E) &= \mathbb{E}[z] \\ &= \frac{2\pi\lambda_E 10^{\tilde{\sigma}_S^2}}{\mathcal{L}_0} \int_0^\infty \frac{r}{1+r^\alpha} dr \\ &= \frac{2\pi^2\lambda_E 10^{\tilde{\sigma}_S^2}}{\mathcal{L}_0\alpha} \csc\left(\frac{2\pi}{\alpha}\right)\end{aligned}\quad (17)$$

which completes the proof. $\blacksquare$

From Theorem 1, we deduce the following corollary.

Corollary 1: The average power of the channel estimation error in (9) is given by

$$\mathbb{E}[|e_{mni}|^2] = \xi(\lambda_E) + \frac{\sigma_N^2}{\rho P}.\quad (18)$$

Proof: The proof follows directly from applying Theorem 1 to (9). $\blacksquare$

Remark 1: We notice that the average channel estimation error power depends linearly on the intensity of the Poisson process of the interferers that, in turn, depends on the number

of available orthogonal pilots. We also remark that from (12) it follows directly that $\xi(\lambda_E) = (1/P)\xi(\lambda)$.

Next, we exploit the result of Theorem 1 to formulate a performance metric that enables us to discern the impact on the system performance of forming the cluster $\mathcal{C}_i$ with a given set of APs. This metric is designed to shed light on the variations in the performance of joint detection and decoding via AP signal combining for the packet transmitted by one of the devices, especially in response to changes occurring under distinct network load conditions.

Let us reformulate the received signal samples in (2) by isolating the signal transmitted by user i , as

$$\mathbf{y}_{mn,D} = \sqrt{\rho} \mathbf{x}(\mathbf{W}_i) h_{mni} + \sqrt{\rho} \sum_{j \in \mathcal{I}_i} \mathbf{x}(\mathbf{W}_j) h_{mnj} + \mathbf{v}_{mn}. \quad (19)$$

Let us also consider the estimated and actual channel gains at all the APs in the cluster stacked into two vectors, respectively, $\hat{\mathbf{q}}_i$ and $\mathbf{q}_i = [\mathbf{h}_{i1}^T, \dots, \mathbf{h}_{iC_i}^T]^T$, where $\mathbf{h}_{mi}^T = [h_{m1i}, \dots, h_{mN_i}]$ with $m \in \mathcal{C}_i$. Thus, the estimated channel gains can be expressed in a compact form as

$$\hat{\mathbf{q}}_i = \mathbf{q}_i + \tilde{\mathbf{e}}_i \quad (20)$$

where the channel estimation errors at all the APs in the cluster can be stacked in vector $\tilde{\mathbf{e}}_i = [\mathbf{e}_{i1}^T, \dots, \mathbf{e}_{iC_i}^T]^T$, where $\mathbf{e}_{mi}^T = [e_{m1i}, \dots, e_{mN_i}]$ with $m \in \mathcal{C}_i$. Following the cooperative detection and decoding scheme, the signals of all the $\mathcal{C}_i$ APs forming cluster $\mathcal{C}_i$ are combined via MRC as in (7). The resulting signal vector, without normalization term, becomes

$$\begin{aligned} \mathbf{v}_{i,N} &= \hat{\mathbf{q}}_i^H \tilde{\mathbf{Y}}_{i,D} \\ &= \sqrt{\rho} \mathbf{x}(\mathbf{W}_i) \hat{\mathbf{q}}_i^H \mathbf{q}_i + \sqrt{\rho} \sum_{j \in \mathcal{I}_i} \mathbf{x}(\mathbf{W}_j) \hat{\mathbf{q}}_i^H \mathbf{q}_j + \hat{\mathbf{q}}_i^H \tilde{\mathbf{V}}_i \\ &= \sqrt{\rho} \mathbf{x}(\mathbf{W}_i) \mathbf{q}_i^H \mathbf{q}_i + \sqrt{\rho} \mathbf{x}(\mathbf{W}_i) \tilde{\mathbf{e}}_i^H \mathbf{q}_i + \hat{\mathbf{q}}_i^H \tilde{\mathbf{V}}_i \\ &\quad + \sqrt{\rho} \sum_{j \in \mathcal{I}_i} \mathbf{x}(\mathbf{W}_j) \hat{\mathbf{q}}_i^H \mathbf{q}_j \end{aligned} \quad (21)$$

where we recall that $\tilde{\mathbf{Y}}_{i,D}$ is the aggregate of the payload signals received by all the antennas of the APs in cluster $\mathcal{C}_i$, and $\tilde{\mathbf{V}}_i = [\mathbf{V}_{i1}^T, \dots, \mathbf{V}_{iC_i}^T]^T$ is the aggregate of the noise samples at the APs, with $\mathbf{V}_m^T = [\mathbf{v}_{m1}^T, \dots, \mathbf{v}_{mN}^T]$ and $m \in \mathcal{C}_i$.

Our focus now lies in assessing a performance metric that establishes a connection between the power of the reference signal received from device i and the power of the unwanted contributions after MRC, due to noise, interference and channel estimation error. This evaluation will enable us to gain insights into how the system performance is affected by choice of the subset of the APs to be incorporated into the cluster. Hence, we define the performance metric $\gamma_i(\mathcal{C}_i)$ as the ratio between the average received power from user i and the average interference and noise power after MRC of the signals received from the APs in cluster $\mathcal{C}_i$, i.e.,

$$\gamma_i(\mathcal{C}_i) = \frac{\rho \mathbb{E} \left[|\mathbf{q}_i^H \mathbf{q}_i|^2 \right]}{\rho \mathbb{E} \left[\left| \sum_{j \in \mathcal{I}_i} \hat{\mathbf{q}}_i^H \mathbf{q}_j \right|^2 \right] + \rho \mathbb{E} \left[|\tilde{\mathbf{e}}_i^H \mathbf{q}_i|^2 \right] + \sigma_N^2 \mathbb{E} [\|\hat{\mathbf{q}}_i\|^2]} \quad (22)$$

where the expectation is taken over all the sources of randomness, by keeping fixed the positions of the APs in the cluster $\mathcal{C}_i$ with respect to the reference device i . In our analysis, we assume that the elements of $\tilde{\mathbf{e}}_i$ and $\mathbf{q}_j$ are uncorrelated for $j \in \mathcal{I}_i$. In practice, the correlation between the elements of the two vectors is low and becomes negligible when the number of pilots is sufficiently high.

For the sake of conciseness, let us define the sum of the first and second-order noncentral moments of the large-scale fading coefficients between the APs in the cluster and the reference device i as

$$\mu_1^{(i)} = \sum_{m \in \mathcal{C}_i} \mathbb{E}[\beta_{mi}] \quad (23)$$

$$\begin{aligned} \mu_2^{(i)} &= \mathbb{E} \left[\sum_{m \in \mathcal{C}_i} \beta_{mi}^2 + N \left(\sum_{m \in \mathcal{C}_i} \beta_{mi} \right)^2 \right] \\ &= N \left(\mu_1^{(i)} \right)^2 + \sum_{m \in \mathcal{C}_i} \mathbb{V}[\beta_{mi}] \end{aligned} \quad (24)$$

respectively. Let us also define the quantity $\Upsilon(\lambda, i)$ as

$$\Upsilon(\lambda, i) = \frac{C_i}{P} \left(\xi(\lambda) + \frac{\sigma_N^2}{\rho} \right) \quad (25)$$

which is directly related to the power of the channel estimation error (see the proof of Theorem 2). We proceed to obtain an expression for $\gamma_i(\mathcal{C}_i)$ by exploiting the result of Theorem 1 and by taking the expectation with respect to fast fading and noise. The terms $\mu_1^{(i)}$ and $\mu_2^{(i)}$ can be estimated using Monte Carlo methods.

Theorem 2: The ratio between the average received power from user i and the average interference and noise power after MRC of the signals received by the APs in cluster $\mathcal{C}_i$ can be approximated as

$$\gamma_i(\mathcal{C}_i) \approx \frac{\mu_2^{(i)}}{\xi(\lambda) \left(\Upsilon(\lambda, i) + \mu_1^{(i)} \right) + \frac{1}{C_i} \Upsilon(\lambda, i) \mu_1^{(i)} + \frac{\sigma_N^2}{\rho} \left(\Upsilon(\lambda, i) + \mu_1^{(i)} \right)}. \quad (26)$$

Proof: Let us start by expressing $\mathbb{E}[|\mathbf{q}_i^H \mathbf{q}_i|^2]$. We have

$$\begin{aligned} \mathbb{E} \left[|\mathbf{q}_i^H \mathbf{q}_i|^2 \right] &= \mathbb{E} \left[\left| \sum_{m \in \mathcal{C}_i} \sum_{n=1}^N h_{mni}^H h_{mni} \right|^2 \right] \\ &= \mathbb{E} \left[\left(\sum_{m \in \mathcal{C}_i} \sum_{n=1}^N |h_{mni}|^2 \right)^2 \right] \\ &= N \mathbb{E} \left[\sum_{m \in \mathcal{C}_i} \beta_{mi}^2 + N \left(\sum_{m \in \mathcal{C}_i} \beta_{mi} \right)^2 \right] \\ &= N \mu_2^{(i)}. \end{aligned} \quad (27)$$

Next, we can derive an approximation of the first term of the denominator of (22) by applying Theorem 1 under the assumption of uncorrelation between the elements of $\tilde{\mathbf{e}}_i$ and $\mathbf{q}_j$

for $j \in \mathcal{I}_i$. Specifically, we can write

$$\begin{aligned}
\mathbb{E} \left[\left| \sum_{j \in \mathcal{I}_i} \hat{\mathbf{q}}_i^H \mathbf{q}_j \right|^2 \right] &= \sum_{j \in \mathcal{I}_i} \sum_{m \in \mathcal{C}_i} \sum_{n=1}^N \mathbb{E} \left[|\hat{h}_{mni}|^2 |h_{mnj}|^2 \right] \\
&\approx \sum_{m \in \mathcal{C}_i} \sum_{n=1}^N \mathbb{E} \left[|\hat{h}_{mni}|^2 \right] \mathbb{E} \left[\left| \sum_{j \in \mathcal{I}_i} h_{mnj} \right|^2 \right] \\
&= \xi(\lambda) \sum_{m \in \mathcal{C}_i} \sum_{n=1}^N \mathbb{E} \left[|\hat{h}_{mni}|^2 \right] \\
&= N\xi(\lambda) \left(\frac{C_i}{P} \left(\xi(\lambda) + \frac{\sigma_N^2}{\rho} \right) + \sum_{m \in \mathcal{C}_i} \mathbb{E}[\beta_{mi}] \right) \\
&= N\xi(\lambda) \left(\Upsilon(\lambda, i) + \mu_1^{(i)} \right). \tag{28}
\end{aligned}$$

The second term in the denominator of (22) can be manipulated as

$$\begin{aligned}
\mathbb{E} \left[\left| \hat{\mathbf{e}}_i^H \mathbf{q}_i \right|^2 \right] &= \mathbb{E} \left[\left| \sum_{m \in \mathcal{C}_i} \sum_{n=1}^N e_{mni}^H h_{mni} \right|^2 \right] \\
&= \sum_{m \in \mathcal{C}_i} \sum_{n=1}^N \mathbb{E} \left[|e_{mni}|^2 \right] \mathbb{E} \left[|h_{mni}|^2 \right] \\
&= \frac{N}{P} \left(\xi(\lambda) + \frac{\sigma_N^2}{\rho} \right) \sum_{m \in \mathcal{C}_i} \mathbb{E}[\beta_{mi}] \\
&= \frac{N}{C_i} \Upsilon(\lambda, i) \mu_1^{(i)} \tag{29}
\end{aligned}$$

where Theorem 1 is used to calculate the expectation of the channel estimate error power. Moreover, we can express the last term in the denominator of (22) as

$$\begin{aligned}
\mathbb{E} \left[\|\hat{\mathbf{q}}_i\|^2 \right] &= N \left(\sum_{m \in \mathcal{C}_i} \mathbb{E}[\beta_{mi}] + \frac{C_i}{P} \left(\xi(\lambda) + \frac{\sigma_N^2}{\rho} \right) \right) \\
&= N \left(\mu_1^{(i)} + \Upsilon(\lambda, i) \right). \tag{30}
\end{aligned}$$

Substituting (27)–(30) into (22) yields (26), completing the proof. ■

The insights provided by Theorem 2 aid in comprehending the rationale behind selecting which APs to incorporate into the cluster $\mathcal{C}_i$ for the decoding of the message transmitted by user i . The expression in the denominator of (26) highlights that, including APs with unfavorable large-scale fading coefficients and consequently leading to a pronounced pilot contamination effect, reduces the ratio between the average useful signals' power and that of the interference. From (26), it becomes apparent that, as the cluster size C_i grows, the addition of APs located progressively farther from device i results in a saturation of $\mu_2^{(i)}$. The saturation is a consequence of the diminishing contribution of distant APs to the sum. Conversely, the terms multiplied by C_i exhibit a linear increase, leading to a decrease of the performance metric, as further shown in the numerical results. Moreover, further

insights into the cluster construction process can be acquired by examining $\gamma_i(\mathcal{C}_i)$ when pilot contamination is negligible.

Remark 2: When there is a substantial number of orthogonal pilots available, the impact of pilot contamination becomes insignificant, leading to a simplified form of (26), i.e.,

$$\lim_{P \rightarrow \infty} \gamma_i(\mathcal{C}_i) = \frac{\mu_2^{(i)}}{\mu_1^{(i)} \left(\xi(\lambda) + \frac{\sigma_N^2}{\rho} \right)}. \tag{31}$$

From (31), we observe that, when pilot contamination is negligible but the payload signal of device i is still interfered, it is recommended to form large clusters of APs for cooperative detection and decoding. Therefore, it becomes clear that mitigating the impact of pilot contamination, e.g., by increasing the number of available orthogonal pilots, opens up the potential for additional performance improvements through the combination of signals received by different APs.

IV. SCHEMES FOR COOPERATIVE DETECTION, DECODING, AND SIC

Based on the considerations made in Section III, we now address enhancements to the above-described CS processing, that leverage AP cooperation in payload estimation and decoding. We initially examine the performance offered by an idealized algorithm where the groups of APs to be combined to process the uplink signal are selected, for each active device, by a “genie” with full knowledge of the system. This genie-aided strategy is used as a benchmark for two actual schemes, working with no prior information, introduced afterwards. Under specific conditions, the performance of the developed practical strategies can approach the genie-aided one.

A. Genie-Aided Cooperative Detection, Decoding, and SIC

We consider an idealized setting where the CPU is assisted by a genie with full knowledge of the set of active users, their employed pilots, and the aggregate SINR they experience after any linear combining of signals received by any set of cooperating APs. With such an information the CPU is able to operate a per-user processing, instead of the uninformed per-slot and per-pilot processing implemented in CS and other practical schemes; in particular, the CPU can build optimal AP clusters, each one devoted to the decoding of the packet transmitted by a specific active device. The channel gains between the APs and user i , for each antenna n , are obtained through ML estimation over the pilot $\mathbf{p}(\mathbf{W}_i)$ used by user i , which is known by the genie, as in (8). The channel gains are evaluated for all $m \in \mathcal{C}_i$. Subsequently, MRC is used to combine the signals received by all the APs in $\mathcal{C}_i$ as in (7). Then, decoding of the user's message is attempted from vector $\mathbf{v}_i \in \mathbb{C}^{1 \times N_D}$.

We can define the aggregated SINR for user i resulting from the combining of all the signals received by the APs in the cluster $\mathcal{C}_i$, as

$$\text{SINR}(i, \mathcal{C}_i) = \frac{\rho |\hat{\mathbf{q}}_i^H \mathbf{q}_i|^2}{\rho \sum_{k \in \mathcal{I}_i} |\hat{\mathbf{q}}_i^H \mathbf{q}_k|^2 + \sigma_N^2 \|\hat{\mathbf{q}}_i\|^2} \tag{32}$$

where $\mathbf{q}_i$ is the stacked vector of all channel gains h_{mi} between APs m in the cluster $\mathcal{C}_i$ and user i . The genie-aided scheme aims to find the smallest cluster of APs, $\mathcal{C}_i$, which can successfully decode the packet of user i , maximizing $\text{SINR}(i, \mathcal{C}_i)$. It exploits the genie's knowledge of SINR obtained after combining the signals received by any set of APs. The cooperative detection and decoding algorithms are executed by the CPU only after the system has performed the CS processing scheme. This cascading operation yields two main benefits: 1) interference cancelation performed by the CS algorithm reduces the pilot contamination, thereby enhancing the performance of payload detection and estimation through AP combining and 2) the CS scheme proves to be less burdensome in terms of data exchanged over the fronthaul links. Hence, AP combining schemes are employed only when CS cannot successfully decode messages from all users. The genie-aided cooperative detection, decoding, and SIC (CDS-G) scheme is still organized into two subsequent phases.

1) *User-By-User Cooperative Decoding (UCD)*: The CPU iteratively builds clusters of APs that attempt cooperative decoding of the user's messages, and stores all the decoded messages in a list $\mathcal{L}$. Initially, the algorithm foresees a cluster composed by a single AP ($\mathcal{C}_i = 1$), which is the one having the largest SINR for the considered user. If the decoding fails, the cardinality of the set $\mathcal{C}_i$ is increased by 1 by adding the AP that leads to the maximum $\text{SINR}(i, \mathcal{C}_i)$.

2) *SIC*: the CPU performs SIC in the slots in which the decoded messages were transmitted and reattempts cooperative decoding of the messages not yet decoded. When a new message is decoded, it is added to the list $\mathcal{L}$.

The complete pseudocode of the genie-based algorithm is given in Algorithm 3, where $\mathcal{W}$ is the set of devices whose message remains undecoded, $[N_s]_i(r)$ is the slot in which replica r of message i has been transmitted, and $C_{\max}$ is the maximum cardinality allowed for the set of cooperating APs, which corresponds also to the maximum number of attempts to decode one message. Clearly, this idealized AP cooperation scheme is not feasible in practice, as it takes advantage of the knowledge of a genie, having important information about the network users at its disposal. In the following we propose two practical schemes for AP cooperation that do not require any prior information about the network users.

B. Distance-Based Cooperative Detection and Decoding

The main idea behind AP combining is to identify sets of APs that can jointly perform the decoding of the message transmitted by a user. However, in the grant-free access schemes the set of active devices, and the pilots used by each device, are not known. Hence, in every slot of the frame and for each pilot, the CPU is required to blindly process the signals received by all APs to detect potential clusters of receivers that should collaborate to jointly decode the messages of active users. Note that, since channel state information is unknown to the receivers, it is also hard to estimate the SINR after MRC and to exploit it to build such clusters. In the following, we propose a distance-based cooperative detection, decoding, and SIC (CDS-A) algorithm

Algorithm 3: Pseudo-code for CDS-G Scheme

Phase 1 - User-by-User Cooperative Decoding (UCD)

```

1  forall  $i \in \mathcal{W}$  do
2     $r \leftarrow 1$ 
3    do
4      Process slot  $[N_s]_i(r)$ 
5       $\mathcal{C}_i \leftarrow \emptyset$ 
6       $u \leftarrow \arg \max_{m \notin \mathcal{C}_i} \text{SINR}(i, \{m\})$ 
7       $\mathcal{C}_i \leftarrow \mathcal{C}_i \cup \{u\}$ 
8       $C \leftarrow 1$ 
9      do
10        $\hat{\mathbf{h}}_{mi} \leftarrow \mathbf{Y}_{m, \text{PP}}^H(\mathbf{W}_i) / \|\mathbf{p}(\mathbf{W}_i)\|^2, \forall m \in \mathcal{C}_i$ 
11        $\mathbf{v}_i \leftarrow \hat{\mathbf{q}}_i^H \tilde{\mathbf{Y}}_{i, \text{D}} / \|\hat{\mathbf{q}}_i\|^2$ 
12       Attempt decoding  $\mathbf{v}_i$ 
13       if decoding successful then
14          $\mathcal{L} \leftarrow \{\mathcal{L}, \mathbf{W}_i\}$ 
15         Remove  $i$  from  $\mathcal{W}$ 
16       else
17          $u \leftarrow \arg \max_{m \notin \mathcal{C}_i} \text{SINR}(i, \mathcal{C}_i \cup \{m\})$ 
18          $\mathcal{C}_i \leftarrow \mathcal{C}_i \cup \{u\}$ 
19       end
20        $C \leftarrow |\mathcal{C}_i|$ 
21     while  $C \leq C_{\max}$  and  $\mathbf{W}_i$  not decoded
22     if  $\mathbf{W}_i$  not decoded then
23        $r \leftarrow r + 1$ 
24     end
25   while  $\mathbf{W}_i$  not decoded and  $r \leq R$ 
26 end

```

Phase 2 - SIC

```

27 forall  $l \in \mathcal{L}$  do
28   Execute DIC( $[N_s]_l$ )
29   Execute UCD( $\mathcal{W}$ )
30 end

```

that does not require any a priori insight into the active users in the network. The cooperation mechanism is based on the creation of M clusters, which depend solely on the geometry of the APs, each one having as centroid a different AP of the network.

We say that AP k is a neighbor of AP m if the distance r_{mk} is less than or equal to a threshold $D_{\max}$, i.e., $r_{mk} \leq D_{\max}$. Then, the cluster $\mathcal{C}_m = \{t_1, \dots, t_{C_i}\}$, $C_i = |\mathcal{C}_m|$, which has AP m as its centroid, is composed by all the neighbors of m , with m included. It is important to note that in this case, differently from the genie-aided algorithm, the notation $\mathcal{C}_m$ identifies a cluster that has AP m as the centroid. The signal received by each AP in $\mathcal{C}_m$ is projected along the j th pilot to perform channel estimation as in (4). The estimate of all channel gains can be stacked in a single vector $\hat{\mathbf{q}}_m(j) = [\hat{\mathbf{h}}_{t_1}(j)^T, \dots, \hat{\mathbf{h}}_{t_{C_i}}(j)^T]^T$, where $\hat{\mathbf{h}}_m(j)^T = [\hat{h}_{m1}(j), \dots, \hat{h}_{mN}(j)]$. Then, MRC is performed as

$$\mathbf{v}_m(j) = \frac{\hat{\mathbf{q}}_m(j)^H \tilde{\mathbf{Y}}_{m, \text{D}}}{\|\hat{\mathbf{q}}_m(j)\|^2} \quad (33)$$

and finally decoding is attempted from $\mathbf{v}_m(j) \in \mathbb{C}^{1 \times N_D}$.

The proposed algorithm is executed after CS scheme and retains the structure of CDS-G, but with a few variations described hereafter. First, distance-based clusters are built based on the network geometry. Then, the CPU attempts

Algorithm 4: Pseudo-code for CDS-A Scheme

```

1 for  $m \leftarrow 1$  to  $M$  do
2    $C_m \leftarrow$  all the APs  $j : r_{m,j} \leq D_{\max}$ 
3 end

```

$\left. \begin{array}{l} \text{RD}(C_1, \dots, C_M; D_{\max}) \end{array} \right\}$

Phase 1 - Slot-by-slot Cooperative Decoding (SCD)

```

4 forall  $s \in [N_s]$  do
5   Process slot  $s$ 
6   for  $m \leftarrow 1$  to  $M$  do
7     for  $j \leftarrow 1$  to  $P$  do
8        $C \leftarrow |C_m|$ 
9        $\hat{h}_k(j) \leftarrow Y_{k,p} p_j^H / \|p_j^H\|^2, \forall k \in C_m$ 
10       $v_m(j) \leftarrow \hat{q}_m(j)^H \tilde{Y}_{m,D} / \|\hat{q}_m(j)\|^2$ 
11      Attempt decoding  $v_m(j)$ 
12      if decoding successful then
13         $\mathcal{L} \leftarrow \{\mathcal{L}, W\}$ 
14      end
15    end
16  end
17 end

```

$\left. \begin{array}{l} \text{DD}(C_{m,j}) \\ \text{SCD}([N_s]) \end{array} \right\}$

Phase 2 - SIC

```

18 forall  $l \in \mathcal{L}$  do
19   Execute DIC( $[N_s]_l$ )
20   Execute SCD( $[N_s]_l$ )
21 end

```

decoding of messages, for each slot and for each available pilot, using the clusters of APs, storing the decoded messages in $\mathcal{L}$. Subsequently, the information about the decoded messages is used to perform interference cancellation as in phase 2 of CDS-G, and decoding is reattempted by the CPU using the clusters of APs. The complete pseudocode of the proposed CDS-A algorithm is given in Algorithm 4.

C. Correlation-Based Cooperative Detection and Decoding

Although the CDS-A algorithm does not require any a priori information about the active users, its cluster construction process does not make full exploitation of the different amounts of useful and interference powers received by the APs. For this reason, aiming at associating only those APs that receive signals from the same device with favorable SINR, we propose an iterative algorithm for cluster construction based on the similarity of signals received by the APs. At each time slot and for each available pilot the CPU performs the following operations. As in CDS-A algorithm, the CPU creates M clusters, each one having as centroid a different AP of the network. As in CDS-G scheme, the size of each cluster is increased iteratively, until a maximum size $C_{\max}$ is reached. At each iteration, a new AP is added to each cluster as follows. Each AP individually estimates its vector of channel gains using pilot j as in (4), and performs MRC as in (5). The AP a that is added to the cluster C_m having AP m as centroid is

$$a = \arg \max_{k \in \tilde{C}_m \setminus C_m} |z_m(j) z_k^H(j)| \quad (34)$$

where $\tilde{C}_m$ is formed by the APs k for which $r_{mk} \leq D_{\max}$, and the amplitude of the correlation between the signals at the APs is evaluated after performing individual MRC. As for

Algorithm 5: Pseudo-code for CDS-B Scheme

```

1 Execute  $\text{RD}(\tilde{C}_1, \dots, \tilde{C}_M; D_{\max})$ 

```

Phase 1 - Correlation-based SCD (CSCD)

```

2 forall  $s \in [N_s]$  do
3   Process slot  $s$ 
4   for  $m \leftarrow 1$  to  $M$  do
5     forall  $j \leftarrow 1$  to  $P$  do
6        $C_m \leftarrow \{m\}$ 
7       for  $c \leftarrow 2$  to  $C_{\max}$  do
8          $a \leftarrow \arg \max_{k \in \tilde{C}_m \setminus C_m} |z_m(j) z_k^H(j)|$ 
9          $C_m \leftarrow C_m \cup \{a\}$ 
10        Execute  $\text{DD}(C_m, j)$ 
11      end
12    end
13  end
14 end

```

$\left. \begin{array}{l} \text{CSCD}([N_s]) \end{array} \right\}$

Phase 2 - SIC

```

15 forall  $l \in \mathcal{L}$  do
16   Execute  $\text{DIC}([N_s]_l)$ 
17   Execute  $\text{CSCD}([N_s]_l)$ 
18 end

```

CDS-G and CDS-A, the signals received by the APs belonging to the same cluster are stacked, MRC is performed, and then decoding is attempted. The pseudocode of the correlation-based cooperative detection, decoding, and SIC (CDS-B) scheme is given in Algorithm 5.

V. NUMERICAL RESULTS

In this section, we initially assess the validity of the developed analytical model in a controlled scenario and employ it to investigate how the performance metric of the AP signal combining schemes, γ_i , varies across different cluster compositions and under various interference conditions. We also show through a practical example the relationship between γ_i and the statistic of the SINR after MRC of signals received by the APs in the cluster. Subsequently, we assess the performance of the genie-aided and practical cooperative detection and decoding schemes (i.e., CDS-G, CDS-A, and CDS-B) proposed in Section IV over actual scenarios of interest for CF-mMIMO systems. This evaluation is conducted through extensive simulations; the performance is compared with that of the CS and NC schemes.

A. AP Cooperation Performance Metric

We consider a circular area of radius $L = 300$ m with a reference device i located in the center, $M = 20$ single-antenna APs randomly positioned with uniform distribution in the annulus with inner and outer radii measuring 3 m and 12 m, respectively, centered in the reference device's position, and a set of active interferers distributed according to a PPP with density $\lambda = 0.015$ interfering devices per m^2 and per slot. We consider a generic slot of the CRA scheme detailed in Section II, in which we assume that all the interferers and the reference device transmit a packet replica picking a random pilot among the P available ones. The corresponding density

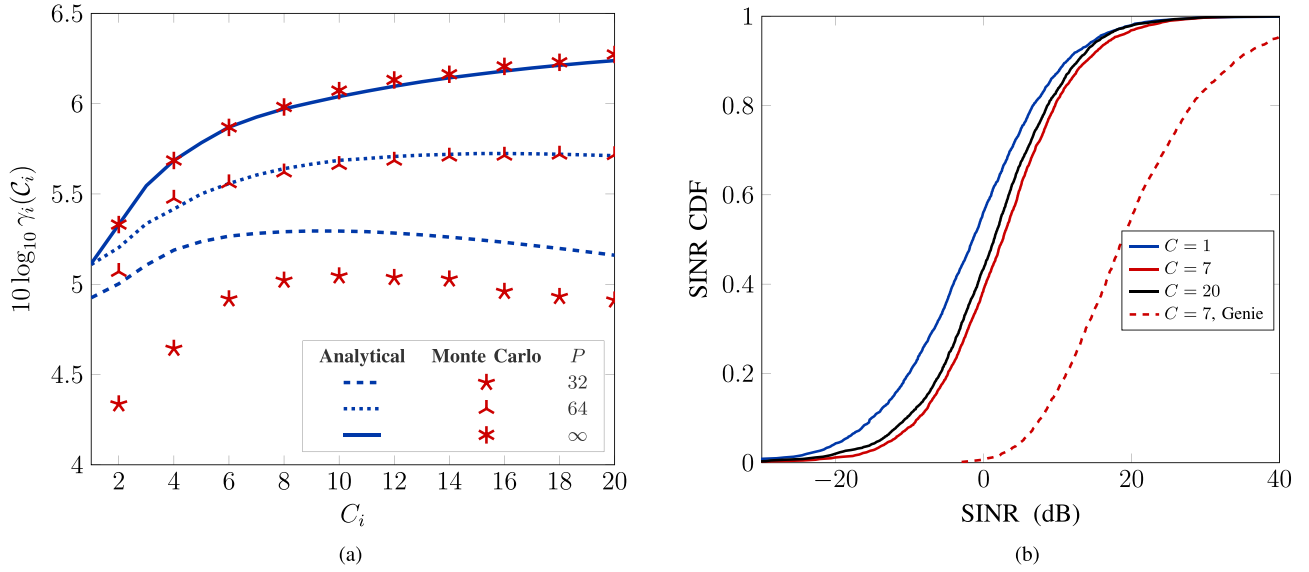


Fig. 3. (a) Performance metric $\gamma(C_i)$ versus the number of APs in cluster C_i , with different numbers of orthogonal pilots, P (i.e., under different pilot contamination regimes). (b) CDF of the SINR under different cluster compositions with $P = 32$.

of the devices adopting the same pilot as the reference device is $\lambda_E = \lambda/P$. The path-loss model is the one defined in (10), with $\mathcal{L}_0 = 40$ dB and $\alpha = 3$. The ratio between the transmit power of each device and the noise power is $\rho/\sigma_N^2 = 130$ dB, and the shadowing coefficient is $\sigma_S = 4$ dB. We compare the analytical expression in (26) with the outcomes of a Monte Carlo simulator.

Fig. 3(a) depicts $\gamma(C_i)$ versus the number of APs in cluster C_i under different pilot contamination regimes. The figure shows that the analytical expression matches well the outcomes of the Monte Carlo simulations when the number of orthogonal pilots is sufficiently large, i.e., in a low-pilot contamination regime. This stems from the assumption that the elements of $\tilde{\mathbf{e}}_i$ and $\mathbf{q}_j$ are uncorrelated for $j \in \mathcal{I}_i$. In practice, the correlation between these vectors' elements is minimal and becomes negligible when the number of available pilots is sufficiently large. However, when the number of orthogonal pilots is limited, this correlation introduces a small discrepancy between the analytical model and Monte Carlo simulations. This discrepancy is not significant, as the trend of γ_i remains consistent, allowing us to draw reliable conclusions about the choice the optimal cluster size for cooperative detection and decoding of the message of device i .

Moreover, Fig. 3(a) shows that when only a small number of orthogonal pilots is available, the impact of pilot contamination induces a degradation in performance when additional APs that are far apart from the target device are incorporated into the cluster. In fact, with $P = 32$ the maximum $\gamma(C_i)$ is obtained, by including in the cluster solely the $C = 7$ nearest APs, suggesting that optimal performance is achieved with small clusters or even individual APs when P is small. Conversely, Fig. 3(a) shows that when the effect of pilot contamination is mild (i.e., $P = 64$), constructing larger clusters opens up the possibility of increasing the system performance. This effect is even more pronounced when the number of available orthogonal pilots becomes large (i.e.,

$P \rightarrow \infty$), making the effect of pilot contamination negligible. In such conditions, adding all the 20 APs to the cluster improves the system performance.

This effect is also illustrated in Fig. 3(b), which presents the CDF of the actual SINR obtained via Monte Carlo simulations for various cluster compositions considered in Fig. 3(a), with $P = 32$ orthogonal pilots. Fig. 3(b) shows an improvement in the SINR CDF when the signal received by user i is processed by a cluster of $C_i = 7$ APs compared to a single AP. However, incorporating all $M = 20$ APs in the cluster leads to a degradation in the CDF. This behavior is consistent with the metric γ_i in Fig. 3(a), where the curve for $P = 32$ peaks around $C_i = 7$.

Additionally, the dashed curve in Fig. 3(b) represents the CDF of the SINR obtained by forming a cluster of 7 APs using the genie-aided algorithm CDS-G. This demonstrates that the SINR statistics can be significantly improved if the cluster is not only limited in size, but also adaptively constructed based on the instantaneous achievable SINR.

B. Cooperative Detection and Decoding Performance

Monte Carlo simulations were run to assess the performance of the genie-aided and the practical cooperation schemes, in terms of packet loss rate (PLR), P_L , versus the number of simultaneously active devices per frame, K . We remark that while a simple simulation scenario was considered for the validation of the analytical derivation in Section V-A, an actual CF-mMIMO system with the following specifications is used to assess the performance of the AP cooperation schemes. We considered a square area of $400 \text{ m} \times 400 \text{ m}$ with $M = 256$ single-antenna APs, connected to a single CPU and evenly distributed on a regular grid (Fig. 1). Each AP is assumed to have a height of 10 m. To limit border effects, devices are assumed to wake up in a square area of $300 \text{ m} \times 300 \text{ m}$ centered on the AP grid. The path-loss between the i th device and the

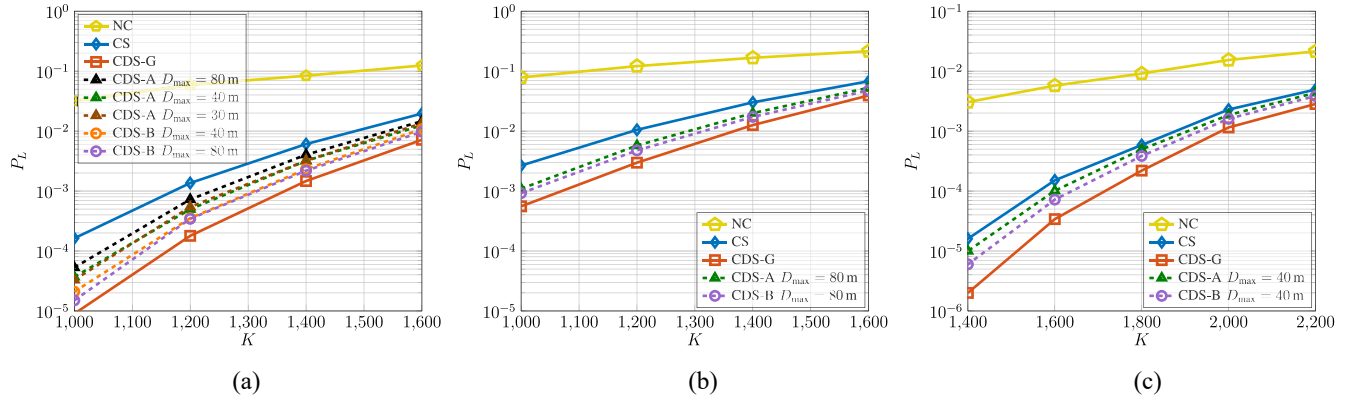


Fig. 4. P_L of the proposed processing schemes versus the number of active devices, K , with $\rho/\sigma_N^2 = 100$ dB and $P = 12$. (a) $M = 256$ and $N = 1$. (b) $M = 64$ and $N = 4$. (c) $M = 256$ and $N = 4$.

m th AP is modeled as

$$PL_{mi}[\text{dB}] = 30.5 + 36.7 \log_{10}(r_{mi}[\text{m}]) \quad (35)$$

where $r_{mi}[\text{m}]$ is the distance between active device i and AP m in meters. The shadowing intensity has been set to $\sigma_S = 6$ dB. As pointed out by several authors (e.g., [10]) this propagation model matches well the 3GPP Urban Microcell model in the sub-6 GHz band, in NLOS conditions and for AP antennas at a height of 10 m. The UE transmit power ranges from -20 to 10 dBm; it is set to 10 and 0 dBm when $\rho/\sigma_N^2 = 100$ dB and $\rho/\sigma_N^2 = 90$ dB, respectively. The message length of each active device is 421 bits. The message is encoded by a (511 421 10) BCH code, a null bit is appended to the codeword, and finally the encoded bits are mapped onto a QPSK constellation, yielding a data payload of $N_D = 256$ symbols. The payload is appended to a pilot sequence of $N_P = 16$ symbols to form the packet. The number of slots per frame is $N_s = 78$, the number of packet replicas is $R = 2$, and the number of available pilot sequences ranges from $P = 1$ to $P = 16$ depending on the test.

1) P_L Versus K in High-Signal-to-Noise Ratio Regime: In Fig. 4(a), the proposed processing schemes are compared in terms of P_L versus the number of active users per frame, K , for $\rho/\sigma_N^2 = 100$ dB and $N = 1$. In this condition, the limiting factor for system performance is the interference among devices. The genie-aided approach demonstrates that performance improvement via AP signal combining is possible, leading to an increase of around 17% of served users at $P_L = 10^{-3}$ with respect to CS and more than 50% with respect to NC. Furthermore, also CDS-A and CDS-B exhibit a performance enhancement with respect to CS. As expected, the improvement is lower than that of CDS-G, since the practical schemes operate blindly, i.e., lacking prior knowledge about the active devices. Moreover, CDS-B is the practical scheme that yields the best performance, proving to be less sensitive than CDS-A to the choice of $D_{\max}$. This is because the CDS-A scheme forms clusters of cooperating APs based solely on their mutual distances, whereas the CDS-B scheme incorporates an iterative algorithm for cluster construction based on the similarity of the signals received by the APs. Consequently, the CDS-B scheme is more adept at identifying

clusters of APs that experience the highest aggregated SINR.

For comparison, we also simulated a scenario with a lower number of multiantenna APs distributed across the area, while maintaining the same total number of antennas. Fig. 4(b) presents the performance of the proposed schemes with $M = 64$ APs, each equipped with $N = 4$ antennas, and a SNR $\rho/\sigma_N^2 = 100$ dB. The total number of APs antennas in this scenario is 256, as in the setting of Fig. 4(a). Comparing the performance of the CDS-G scheme in Fig. 4(b) with that in Fig. 4(a), it is evident that distributing the antennas throughout the network area results in a significant performance gain due to AP cooperation. However, the performance gap between the practical schemes and the genie-aided one remains approximately the same in both configurations.

Fig. 4(c) shows P_L versus K with $\rho/\sigma_N^2 = 100$ dB and APs equipped with $N = 4$ antennas. In this condition, where interference among devices is the primary limiting factor for system performance, the genie-aided approach reveals that performance enhancement through AP cooperation remains viable. This results in an approximate 8% increase in the number of supported users at $P_L = 10^{-3}$ compared to CS and an approximate 70% increase compared to NC. However, the scalability gain with respect to CS for APs with $N = 4$ antennas in such a high-SNR regime is less than that observed for single-antenna APs since an additional gain for all the schemes is due to multiple antenna processing. Similarly, both CDS-A and CDS-B show improved scalability, supporting around 4% more devices at $P_L = 10^{-3}$ compared to CS. As anticipated, the enhancement is lower than that achieved by CDS-G.

2) P_L Versus K in Low-Signal-to-Noise Ratio Regime: Fig. 5(a) shows P_L versus K , setting $\rho/\sigma_N^2 = 90$ dB and $N = 1$. Also in this low-SNR regime, cooperative detection and decoding achieves significantly better performance than NC and CS. The improvement has to be attributed to the effective exploitation of distributed multiple antennas, enabled by AP combining. Fig. 5(a) reveals how AP combining is a powerful tool to achieve good performance in a low-device transmit power regime, thereby increasing devices' battery lifetime and network energy efficiency. This result is

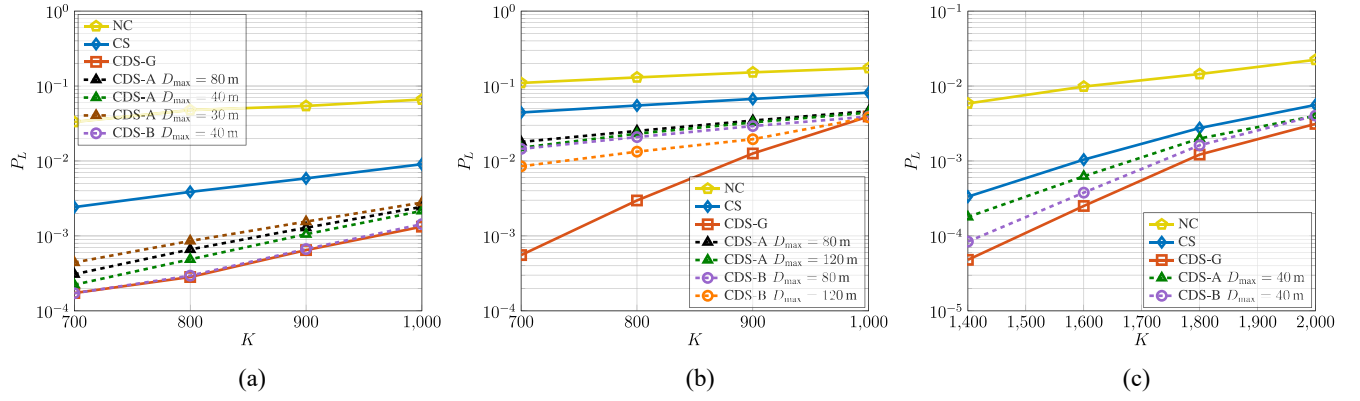


Fig. 5. P_L of the proposed processing schemes versus the number of active devices, K , with $\rho/\sigma_N^2 = 90$ dB and $P = 12$. (a) $M = 256$ and $N = 1$. (b) $M = 64$ and $N = 4$. (c) $M = 256$ and $N = 4$.

especially promising in very low-power IoT applications. Very remarkably, in this regime CDS-B with $D_{\max} = 40$ m performs as good as the CDS-G benchmark.

Fig. 5(b) presents the performance of the proposed schemes with $M = 64$ APs, each equipped with $N = 4$ antennas and $\rho/\sigma_N^2 = 90$ dB. Comparing the performance of CDS-G in Fig. 5(a) with that in Fig. 5(b), it is again evident that distributing the antennas throughout the network area enables a significant performance gain for AP cooperation. Additionally, the gap between the PLR achieved by the practical schemes and that of the genie-aided benchmark appears to be larger in this configuration.

Fig. 5(c) shows P_L versus K with $\rho/\sigma_N^2 = 90$ dB and APs equipped with $N = 4$ antennas. In this low-SNR regime, cooperative detection and decoding still achieves remarkably better performance than CS, even compared with the single-antenna case. The improvement has to be attributed to the effective exploitation of both co-located and distributed multiple antennas, enabled by AP combining. The results in Fig. 5(c) highlight that genie-aided AP combining supports approximately 12.5% more devices at $P_L = 10^{-3}$ compared to CS and more than 50% with respect to NC, underscoring the effectiveness of AP combining in achieving robust performance in a low-device transmit power regime. Notably, in this regime, CDS-B with $D_{\max} = 40$ m remains the best practical cooperation scheme, demonstrating performance remarkably close with that of CDS-G.

3) *Impact of Cooperation on Devices' Energy Consumption:* Fig. 6 illustrates P_L versus the ratio ρ/σ_N^2 in a scenario with $K = 1000$ active users in the frame and $P = 12$ available orthogonal pilots. The figure demonstrates how AP cooperation enhances performance, enabling a reduction in device transmit power without compromising system efficiency. Specifically, achieving $P_L = 10^{-3}$ with CS scheme requires $\rho/\sigma_N^2 = 95$ dB, while CDS-G and CDS-B schemes achieve the same performance at $\rho/\sigma_N^2 = 90$ dB. Note that NC performance is limited by a floor at $P_L > 10^{-2}$. Therefore, the adoption of cooperative detection and decoding schemes can lead to remarkable energy savings on the part of devices, enabling longer battery durations and aligning with the evolution of future mMTC systems designed for low-power applications.

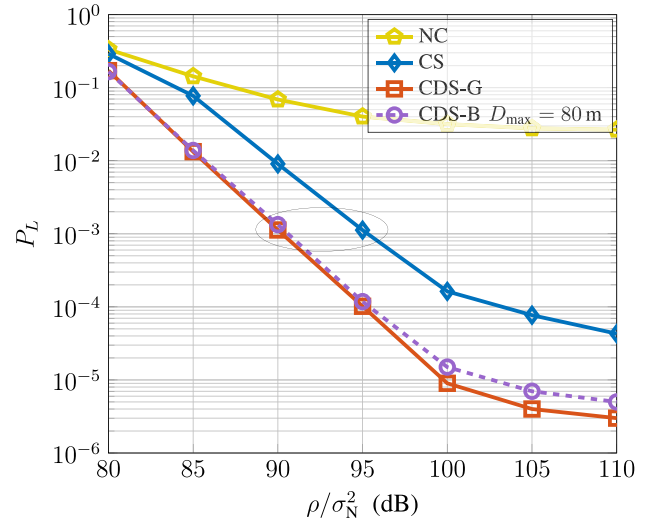


Fig. 6. P_L of the proposed processing schemes versus ρ/σ_N^2 , with $K = 1000$, $P = 12$ orthogonal pilots, and $N = 1$. The performance gap between the proposed schemes is highlighted at $P_L = 10^{-3}$.

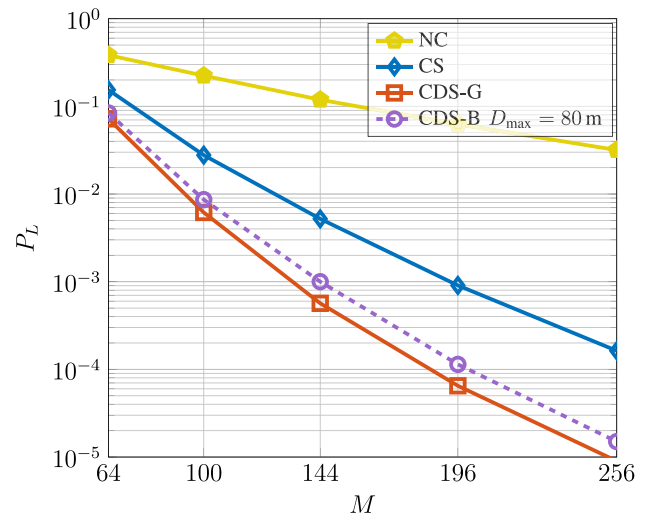


Fig. 7. P_L of the proposed processing schemes versus the number of APs, M , with $\rho/\sigma_N^2 = 100$ dB, $K = 1000$, $P = 12$, and $N = 1$.

4) *Impact of the Number of APs:* Fig. 7 shows P_L versus the number of APs deployed in the area for $K = 1000$ active users in the frame, setting $\rho/\sigma_N^2 = 100$ dB, $P = 12$, and

TABLE I
COMPUTATIONAL COMPLEXITY AND FRONTHAUL LOAD

Algorithm	Computational complexity	Fronthaul load [bit/frame]
NC	$O(MN(N_sP + KR + (K-1)RP))$	kKM
CS	$O(MN(N_sP + KR + (K-1)RP))$	kKM
CDS-G	$O(MNKR + \frac{1}{4}NRC_{\max}K(C_{\max} + 1)(K-1))$	$B(N_P + N_D)NMN_s$
CDS-A	$O(MN(\tilde{C}_{\max}N_sP + KR + (K-1)RP\tilde{C}_{\max}))$	$B(N_P + N_D)NMN_s$
CDS-B	$O(\frac{1}{2}MN(N_sPC_{\max}(C_{\max} + 1) + 2KR + (K-1)RPC_{\max}(C_{\max} + 1)))$	$B(N_P + N_D)NMN_s$

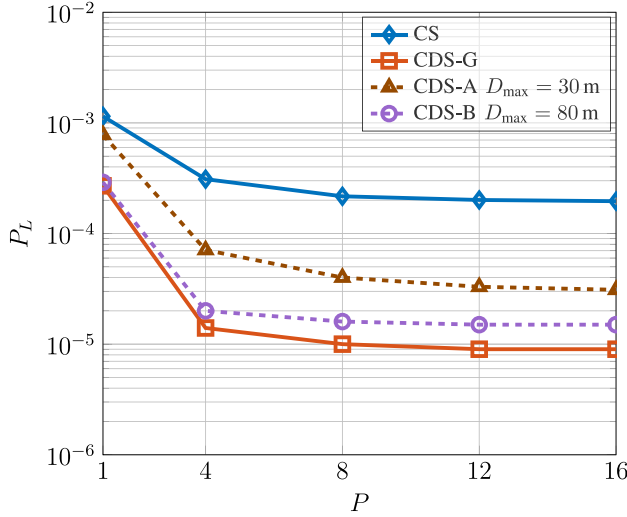


Fig. 8. P_L of the proposed processing schemes versus the number of available orthogonal pilots, P , with $\rho/\sigma_N^2 = 100$ dB, $K = 1000$, and $N = 1$.

$N = 1$. As envisaged, increasing the number of APs leads to a performance improvement since the devices are more likely to be in close proximity to at least one of the APs. Interestingly, the performance gain attainable via AP cooperation increases when there are more APs in the area. Remarkably, $P_L = 10^{-3}$ can be achieved with CDS-G and CDS-B schemes deploying 52 APs less than with CS. This presents an opportunity for network providers to simplify the network infrastructure and decrease the associated deployment costs.

5) *Impact of the Number of Pilots*: Fig. 8 depicts P_L for $K = 1000$ active users as a function of the number of available orthogonal pilots P , setting $\rho/\sigma_N^2 = 100$ dB. As expected, increasing P leads to a performance enhancement since the effect of pilot contamination is reduced. However, from a certain number of available pilots on, P_L saturates due to the remaining effect of payload interference. Furthermore, Fig. 8 shows that increasing P opens up the possibility to obtain a more substantial performance improvement via AP cooperation. Specifically, the performance gains achievable through CDS-G, CDS-A, and CDS-B schemes exhibit an increase when moving from $P = 1$ to $P = 4$.

6) *Computational Complexity and Fronthaul Load*: To conclude the analysis, we evaluate the computational complexity of the proposed processing schemes leveraging Big O notation, that illustrates the efficiency of our algorithms in terms of number of computational steps required for their

execution on a generic machine. We denote by $\tilde{C}_{\max}$ the number of APs involved in a decoding attempt in algorithm CDS-A. Throughout the analysis, the sequence of operations needed to perform channel estimation, payload detection, and message decoding is counted as cN steps, where c is the number of APs involved in the detection and decoding process. For example, $c = 1$ in the NC and CS schemes, $c = \tilde{C}_{\max}$ in CDS-A, and c ranges from 1 to $C_{\max}$ in CDS-G and CDS-B. Also, the sequence of operations needed to perform interference cancellation at one antenna of one AP is counted as 1 step.

We also calculate the amount of fronthaul traffic that the cell-free network architecture has to undergo for each proposed scheme to work. In particular, let us define the fronthaul load as the number of bits that all APs have to send to the CPU for the processing of one entire frame when adopting one of the proposed schemes, and let us denote by B the number of memory bits occupied by a complex number.

The computational complexities and fronthaul loads for all the proposed processing schemes are reported in Table I and detailed in the following.

- 1) *NC*: Phase 1 requires MN_sP steps for all the APs to process all the slots of the frame. Then, each AP subtracts the interference generated by the replicas of the decoded messages, and attempts decoding again in all slots in which the replicas were transmitted. In the worst case, this yields to $MNKR$ steps for interference cancellation and $MN(K-1)RP$ steps for reprocessing the already processed slots. The overall complexity of NC is $O(MN(N_sP + KR + (K-1)RP))$. Considering the fronthaul load, each AP operates independently, transmitting the locally decoded messages to the CPU. In the worst-case scenario, the fronthaul load is kKM bits per frame.
- 2) *CS*: The CS scheme follows the same processing steps of NC, with the sole distinction being that the CPU orchestrates a distributed SIC phase. Consequently, in the worst-case scenario, the computational complexity of the CS scheme is equivalent to that of the NC scheme, i.e., $O(MN(N_sP + KR + (K-1)RP))$. However, we note that the NC scheme is less likely to decode all active users and thereby initiate the corresponding interference cancellations. Consequently, under practical conditions, the NC scheme exhibits lower computational complexity compared to CS, albeit at the cost of reduced system performance. In the worst-case scenario, the

APs transmit a total of kKM bits per frame over the fronthaul.¹

- 3) *CDS-G*: Phase 1 requires at most $(1/2)KRNC_{\max}(C_{\max} + 1)$ steps to process all the users' replicas once. Regarding Phase 2, DIC and UCD procedures require at most $MNKR$ and $(1/4)NRC_{\max}K(C_{\max} + 1)(K - 1)$ steps, respectively. Therefore, the overall complexity of CDS-G scheme is $O(MNKR + (1/4)NRC_{\max}K(C_{\max} + 1)(K - 1))$. It is important to note that, since the CDS-G scheme is a genie-aided benchmark, it involves an exhaustive search for the subsets of APs that maximize the SINR of each user. This results in a computational complexity that scales with the square of the number of users and the square of the maximum AP cluster size, K^2 and $C_{\max}^2$, respectively. Each AP sends $(N_P + N_D)NN_s$ complex numbers to the CPU, resulting in a total fronthaul load of $B(N_P + N_D)NMN_s$ bits per frame.
- 4) *CDS-A*: As this cooperation mechanism relies on the formation of M clusters, determined exclusively by the geometry of the APs, the number of steps required by the CDS-A scheme in the worst-case scenario is similar to that of the NC and CS schemes, with the only difference that the number of APs involved in the decoding attempts is $\tilde{C}_{\max}$. The overall complexity is thus $O(MN(\tilde{C}_{\max}N_sP + KR + (K - 1)RPC_{\max}))$. As for CDS-G, the total fronthaul load is $B(N_P + N_D)NMN_s$ bits per frame.
- 5) *CDS-B*: Phase 1 of the CDS-B scheme requires a maximum of $(1/2)MNN_sPC_{\max}(C_{\max} + 1)$ processing steps. Regarding Phase 2, DIC and CSCD procedures require at most $MNKR$ and $(1/2)MN(K - 1)PRC_{\max}(C_{\max} + 1)$ steps, respectively. Therefore, the overall complexity is $O((1/2)MNN_sPC_{\max}(C_{\max} + 1) + 2KR + (K - 1)RPC_{\max}(C_{\max} + 1))$. As for CDS-G and CDS-A, the total fronthaul load is $B(N_P + N_D)NMN_s$ bits per frame.

As an example, for a system with $M = 256$, $N = 1$, $N_s = 78$, $K = 1000$, $k = 421$ bits, $N_P = 16$ symbols, $N_D = 256$ symbols, assuming that a complex number occupies $B = 128$ bits in memory, the fronthaul load of CS and NC becomes 107.8 Mb/frame, while that of CDS-G, CDS-A, and CDS-B is 695.2 Mb/frame (i.e., around seven times higher).

We also observe that the centralized nature of CDS-A and CDS-B ensures a stable fronthaul load independent of the number of users, making these algorithms scalable and practical in large networks. On the other hand, in more distributed algorithms, such as NC and CS, the fronthaul load depends on K : it initially increases with the density of users, until the system load reaches a critical point; once the system becomes overloaded and the PLR rises to large values, the fronthaul load decreases due to the reduced number of successful decoding events.

¹Note that to execute the distributed SIC the CPU must send up to kKM bits per frame to the APs, thereby generating fronthaul traffic in the opposite direction compared to other processing schemes.

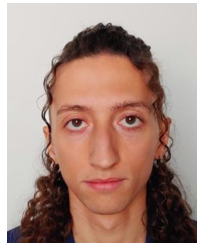
VI. CONCLUSION

In this article, a grant-free MMA scheme based on CRA and specifically designed for CF-mMIMO networks has been presented. Leveraging the distributed nature of the network and the combining of signals received at properly selected APs, the proposed processing enhances the system scalability-versus-reliability tradeoff. Initially, a stochastic geometry-based analytical framework to study the performance of AP cooperation under different cluster compositions has been derived. Then, the potential gain attainable via AP cooperation has been explored using a genie-aided approach. Finally, two practical algorithms for AP selection in a fully grant-free context have been proposed. Numerical results have shown how AP cooperation through AP signal combining and cooperative SIC can bring tangible benefits even without prior information about active users, under both high and low-SNR conditions, closing in some cases the gap to the genie-aided approach. Moreover, it has been shown how the adoption of cooperative detection and decoding schemes can lead to energy saving, enabling longer devices' battery durations and aligning with the evolution of future mMTC systems designed for low-power applications. It should be remarked that AP signal combining comes at a price in terms of increased traffic on the fronthaul, as soft signal samples are forwarded to the CPU, and of increased complexity in the CPU, that is now tasked with deciding which APs should be selected for signal combining, and with decoding messages from combined signals. These aspects represent further directions of investigation.

REFERENCES

- [1] Z. Dawy, W. Saad, A. Ghosh, J. G. Andrews, and E. Yaacoub, "Toward massive machine type cellular communications," *IEEE Wireless Commun.*, vol. 24, no. 1, pp. 120–128, Nov. 2017.
- [2] C. Bockelmann et al., "Massive machine-type communications in 5G: Physical and MAC-layer solutions," *IEEE Commun. Mag.*, vol. 54, no. 9, pp. 59–65, Sep. 2016.
- [3] S. K. Sharma and X. Wang, "Toward massive machine type communications in ultra-dense cellular IoT networks: Current issues and machine learning-assisted solutions," *IEEE Commun. Surveys Tuts.*, vol. 22, no. 1, pp. 426–471, 1st Quart., 2020.
- [4] S. R. Pokhrel, J. Ding, J. Park, O.-S. Park, and J. Choi, "Towards enabling critical mMTC: A review of URLLC within mMTC," *IEEE Access*, vol. 8, pp. 131796–131813, 2020.
- [5] E. Björnson and L. Sanguinetti, "Scalable cell-free massive MIMO systems," *IEEE Trans. Commun.*, vol. 68, no. 7, pp. 4247–4261, Jul. 2020.
- [6] E. Björnson and L. Sanguinetti, "Making cell-free massive MIMO competitive with MMSE processing and centralized implementation," *IEEE Trans. Wireless Commun.*, vol. 19, no. 1, pp. 77–90, Jan. 2020.
- [7] J. Liu, M. Sheng, L. Liu, and J. Li, "Network densification in 5G: From the short-range communications perspective," *IEEE Commun. Mag.*, vol. 55, no. 12, pp. 96–102, Dec. 2017.
- [8] H. Q. Ngo, A. Ashikhmin, H. Yang, E. G. Larsson, and T. L. Marzetta, "Cell-free massive MIMO versus small cells," *IEEE Trans. Wireless Commun.*, vol. 16, no. 3, pp. 1834–1850, Mar. 2017.
- [9] T. Van Chien, E. Björnson, and E. G. Larsson, "Joint power allocation and user association optimization for massive MIMO systems," *IEEE Trans. Wireless Commun.*, vol. 15, no. 9, pp. 6384–6399, Sep. 2016.
- [10] Ö. T. Demir, E. Björnson, and L. Sanguinetti, "Foundations of user-centric cell-free massive MIMO," *Found. Trends Signal Process.*, vol. 14, nos. 3–4, pp. 162–472, 2021.
- [11] T. Van Chien, H. Q. Ngo, S. Chatzinotas, M. Di Renzo, and B. Ottersten, "Reconfigurable intelligent surface-assisted cell-free massive MIMO systems over spatially-correlated channels," *IEEE Trans. Wireless Commun.*, vol. 21, no. 7, pp. 5106–5128, Dec. 2022.
- [12] G. Torcolacci, N. Decarli, and D. Dardari, "Holographic MIMO communications exploiting the orbital angular momentum," *IEEE Open J. Commun. Soc.*, vol. 4, pp. 1452–1469, 2023.

- [13] Z. Wang et al., "A tutorial on extremely large-scale MIMO for 6G: Fundamentals, signal processing, and applications," *IEEE Commun. Surveys Tuts.*, early access, Jan. 2, 2024, doi: [10.1109/COMST.2023.3349276](https://doi.org/10.1109/COMST.2023.3349276).
- [14] E. Paolini, Č. Stefanović, G. Liva, and P. Popovski, "Coded random access: Applying codes on graphs to design random access protocols," *IEEE Commun. Mag.*, vol. 53, no. 6, pp. 144–150, Jun. 2015.
- [15] J. Haghighat and T. M. Duman, "An energy-efficient feedback-aided irregular repetition slotted ALOHA scheme and its asymptotic performance analysis," *IEEE Trans. Wireless Commun.*, vol. 22, no. 12, pp. 9808–9820, Dec. 2023.
- [16] J. H. Sørensen, E. De Carvalho, Č. Stefanović, and P. Popovski, "Coded pilot random access for massive MIMO systems," *IEEE Trans. Wireless Commun.*, vol. 17, no. 12, pp. 8035–8046, Dec. 2018.
- [17] E. Sandgren, A. G. I. Amat, and F. Brännström, "On frame asynchronous coded slotted ALOHA: Asymptotic, finite length, and delay analysis," *IEEE Trans. Commun.*, vol. 65, no. 2, pp. 691–704, Feb. 2017.
- [18] M. Berlioli, G. Cocco, G. Liva, and A. Munari, "Modern random access protocols," *Found. Trends Netw.*, vol. 10, no. 4, pp. 317–446, 2016.
- [19] G. Liva, "Graph-based analysis and optimization of contention resolution diversity slotted ALOHA," *IEEE Trans. Commun.*, vol. 59, no. 2, pp. 477–487, Feb. 2011.
- [20] E. Testi, V. Tralli, and E. Paolini, "Coded random access in cell-free massive MIMO networks with access point signal combining," in *Proc. IEEE Int. Conf. Commun.*, 2024, pp. 1–6.
- [21] L. Valentini, E. Paolini, and M. Chiani, "Massive grant-free access in cell-free massive MIMO networks," in *Proc. 28th Eur. Wireless Conf.*, 2023, pp. 76–81.
- [22] J. Li, H. Wang, P. Zhu, D. Wang, and X. You, "Covariance-based activity detection with orthogonal pilot sequences for cell-free distributed massive MIMO systems," *IEEE Trans. Veh. Technol.*, vol. 72, no. 1, pp. 1307–1312, Jan. 2023.
- [23] U. K. Ganesan, E. Björnson, and E. G. Larsson, "An algorithm for grant-free random access in cell-free massive MIMO," in *Proc. IEEE 21st Int. Workshop Signal Process. Adv. Wireless Commun.*, 2020, pp. 1–5.
- [24] H. Iimori, T. Takahashi, K. Ishibashi, G. T. F. De Abreu, and W. Yu, "Grant-free access via bilinear inference for cell-free MIMO with low-coherence pilots," *IEEE Trans. Wireless Commun.*, vol. 20, no. 11, pp. 7694–7710, Nov. 2021.
- [25] X. Wang, A. Ashikhmin, Z. Dong, and C. Zhai, "Two-stage channel estimation approach for cell-free IoT with massive random access," *IEEE J. Sel. Areas Commun.*, vol. 40, no. 5, pp. 1428–1440, May 2022.
- [26] S. Rao, A. Ashikhmin, and H. Yang, "Internet of Things based on cell-free massive MIMO," in *Proc. 53rd Asilomar Conf. Signals Syst. Comput.*, 2019, pp. 1946–1950.
- [27] S. Chen, J. Zhang, E. Björnson, J. Zhang, and B. Ai, "Structured massive access for scalable cell-free massive MIMO systems," *IEEE J. Sel. Areas Commun.*, vol. 39, no. 4, pp. 1086–1100, Apr. 2021.
- [28] M. Sarker and A. O. Fapojuwo, "Granting massive access by adaptive pilot assignment scheme for scalable cell-free massive MIMO systems," in *Proc. IEEE Veh. Tech. Conf.*, 2021, pp. 1–5.
- [29] M. Rahmani, M. J. Dehghani, P. Xiao, M. Bashar, and M. Debbah, "Multi-agent reinforcement learning-based pilot assignment for cell-free massive MIMO systems," *IEEE Access*, vol. 10, pp. 120492–120502, 2022.
- [30] N. Rajapaksha, K. B. S. Manosha, N. Rajatheva, and M. Latva-Aho, "Deep learning-based power control for cell-free massive MIMO networks," in *Proc. IEEE Int. Conf. Commun.*, 2021, pp. 1–7.
- [31] M. Zaher, Ö. T. Demir, E. Björnson, and M. Petrova, "Learning-based downlink power allocation in cell-free massive MIMO systems," *IEEE Trans. Wireless Commun.*, vol. 22, no. 1, pp. 174–188, Jan. 2023.
- [32] J. Zhang, J. Fan, J. Zhang, D. W. K. Ng, Q. Sun, and B. Ai, "Performance analysis and optimization of NOMA-based cell-free massive MIMO for IoT," *IEEE Internet Things J.*, vol. 9, no. 12, pp. 9625–9639, Jun. 2022.
- [33] E. Paolini, L. Valentini, V. Tralli, and M. Chiani, "Irregular repetition slotted ALOHA in an information-theoretic setting," in *Proc. IEEE Int. Symp. Inf. Theory*, 2022, pp. 3019–3024.
- [34] J. F. C. Kingman, *Poisson Processes*. Oxford, U.K.: Clarendon Press, 1993.
- [35] H. Inaltekin, M. Chiang, H. V. Poor, and S. B. Wicker, "On unbounded path-loss models: Effects of singularity on wireless network performance," *IEEE J. Sel. Areas Commun.*, vol. 27, no. 7, pp. 1078–1092, Sep. 2009.
- [36] R. K. Ganti and M. Haenggi, "Interference and outage in clustered wireless ad hoc networks," *IEEE Trans. Inf. Theory*, vol. 55, no. 9, pp. 4067–4086, Sep. 2009.
- [37] D. Stoyan, W. S. Kendall, and J. Mecke, *Stochastic Geometry and Its Applications*, 2nd ed. Hoboken, NJ, USA: Wiley, 1996.



Enrico Testi (Member, IEEE) received the M.S. degree in electronics and telecommunications engineering for energy (*magna cum laude*) and the Ph.D. degree in electronics, telecommunications, and information technologies engineering from the University of Bologna, Cesena, Italy, in 2018 and 2022, respectively.

He is currently the Junior Assistant Professor with the Department of Electrical, Electronic, and Information Engineering "Guglielmo Marconi", University of Bologna. He is also the Secretary/Treasurer of the IEEE ITSoc Italy Section Chapter. He has been affiliated with the National Laboratory of Wireless Communications of the Italian National Interuniversity Consortium for Telecommunications since 2020. His research interests include artificial intelligence techniques for next-generation wireless networks, massive MIMO, and satellite IoT.



Velio Tralli (Senior Member, IEEE) received the Dr.Ing. degree (*cum laude*) in electronic engineering and the Ph.D. degree in electronic engineering and computer science from the University of Bologna, Cesena, Italy, in 1989 and 1993, respectively.

From 1994 to 1999, he was a Researcher with the National Research Council, CSITE, University of Bologna. In 1999, he joined the Department of Engineering, University of Ferrara, Ferrara, Italy, where he is currently an Associate Professor. He has co-authored more than 150 articles in refereed journals, including IEEE transactions and international conferences. He has participated in several national and European Research Projects, and European Cooperation Programs addressing short-range communications systems, 3G/4G/5G wireless networks, wireless video communications, and on-chip optical wireless networks. His research interests include digital transmission and coding and wireless communications, with emphasis on radio resource optimization, cross-layer design, and multiantenna systems.

Dr. Tralli served as the Co-Chair for the Wireless Communication Symposium of ICC 2006 and the Communication Theory Symposium of ICC 2013. He also serves as an Associate Editor for the *European Transactions on Emerging Technologies*.



Enrico Paolini (Senior Member, IEEE) received the Dr.Ing. degree (*summa cum laude*) in telecommunications engineering and the Ph.D. degree in electrical engineering from the University of Bologna, Cesena, Italy, in 2003 and 2007, respectively.

While working toward the Ph.D. degree, he was a Visiting Research Scholar with the Department of Electrical Engineering, University of Hawai'i at Manoa, Honolulu, HI, USA. He was a Visiting Scientist with the Institute of Communications and Navigation, German Aerospace Center, Cologne, Germany, in 2012 and 2014, under DLR-DAAD fellowships. He is currently an Associate Professor with the Department of Electrical, Electronic, and Information Engineering, University of Bologna. His research interests include digital communication systems, error correcting codes, massive multiple access protocols, and detection and tracking in radar systems.

Dr. Paolini served as a Co-Chair for the ICC 2014, ICC 2015, and ICC 2016 Workshop on Massive Uncoordinated Access Protocols, the VTC 2019-Fall Workshop on Small Data Networks, the 2018 IEEE European School of Information Theory, and the 2020 IEEE Information Theory Workshop (ITW 2020). He served as a TPC Co-Chair for the IEEE GLOBECOM 2022-Communication Theory Symposium and the IEEE GLOBECOM 2019-Communication Theory Symposium. He is the Chair of the ITSoc Italy Section Chapter and the Secretary of the IEEE ComSoc Radio Communications Committee. He was an Editor of IEEE COMMUNICATIONS LETTERS from 2012 to 2015 and the IEEE TRANSACTIONS ON COMMUNICATIONS (in coding and information theory) from 2015 to 2020.