IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received September 21, 2020, accepted October 3, 2020, date of publication October 12, 2020, date of current version October 22, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.3030215

Massive Distributed Antenna Systems: Channel Estimation and Signal Detection

PENG PAN[1], (Senior Member, IEEE), JIATIAN ZHANG[2], (Graduate Student Member, IEEE), AND LIE-LIANG YANG[2], (Fellow, IEEE)

[1]School of Communications Engineering, Hangzhou Dianzi University, Hangzhou 310018, China
[2]School of Electronics and Computer Science, The University of Southampton, Southampton SO17 1BJ, U.K.

Corresponding author: Lie-Liang Yang (lly@ecs.soton.ac.uk)

This work was supported in part by the National Natural Science Foundation of China (NSFC) under Grant U1709220 and Grant 61401130, in part by the Zhejiang Provincial Key Laboratory of Technology on Data's Storage, Transmission, and Security, in part by the Engineering and Physical Sciences Research Council (EPSRC) under Project EP/P034284/1, and in part by the Innovate U.K. under the Knowledge Transfer Partnership under Project KTP011036.

ABSTRACT This article addresses the uplink information transmission in the massive distributed antenna (mDA) systems, where the density of antenna elements (AEs) can be much higher than the density of users. We conceive that such mDA systems are user-centric and provide various advantages for practical operations, and that there are evident challenges in the signal detection, which is very different from that in the conventional cellular systems with co-located antennas at base-stations (BS). Therefore in this article, based on the theory of stochastic geometry, some insightful analysis for the achievable performance of mDA systems is first provided, to show that local detection can be more practical than global detection. Then, by exploiting the sparsity of the channel matrices describing mDA systems, a channel estimator is proposed based on the orthogonal matching pursuit (OMP) algorithm. We will show that the OMP-based channel estimation can not only provide channel state information (CSI), but also assist to select the users to be served by an AE. Furthermore, benefiting from the above-mentioned sparsity, we propose a distributed low-complexity message passing algorithm (MPA)-assisted multiuser detector. Finally, we investigate the performance of the mDA systems employing the proposed detection schemes and with various system settings. Our studies show that the OMP-based channel estimation and MPA-assisted detector are capable of achieving a good trade-off among BER performance, complexity and resource usage.

INDEX TERMS Massive distributed MIMO, channel estimation, sparse signal recovery, orthogonal matching pursuit, multiuser detection, message passing algorithm, stochastic geometry.

I. INTRODUCTION

Due to the explosive growth of mobile traffic, and the continued increase in demand for various types of services, the future generations of mobile communication systems are expected to provide high throughput, high energy-efficiency, low latency, and ultra-reliable communications [1]. In order to meet these expectations, many physical layer technologies, including massive multiple-input multiple-output (MIMO), non-orthogonal multiple access (NOMA), full-duplex, index modulation, etc., have been proposed to meet the requirements of future new air interfaces. On the other side, as we have learned from the evolution of cellular systems, a highly effective way to achieve the above-mentioned expectations is to reduce cell size associated with introducing cooperation or coordination among various types of access points (APs) [2], [3]. In this way, the limited spectrum resources can be reused within small and even tiny sized spaces, making APs very close to their served users. Therefore, future wireless systems may be visualized to be a heterogeneous wireless network with ultra-dense deployment of APs, which might no longer be AP-centric, but be *user-centric* instead.

In this article, we conceive a kind of wireless networks, where APs are densely deployed, with the number of APs much more than the number of users. Furthermore, the APs are connected to a central processing unit (CPU) via backhaul links, which enables cooperation/coordination among APs. In principle, the APs can be viewed as the geographically distributed antenna elements (AEs) of the CPU. Hence, the wireless networks considered constitute a distributed antenna system (DAS) with massive number of AEs, which is referred to as the massive distributed antenna (mDA) system.

The mDA system to be addressed has the similar concepts as the traditional DAS [4], virtual/network MIMO [5]–[7],

The associate editor coordinating the review of this manuscript and approving it for publication was Cong Pu.

or super base station (BS) system [8], [9]. However, owing to the ultra-dense deployment of AEs, the main difference between the considered mDA system and the above-mentioned systems is that mDA system is user-centric, where every user may be surrounded by a number of AEs for communications. In other words, every user is the center of a dynamic cell. Recently, the similar concept has also been proposed in the context of cloud radio access network (C-RAN) [10], [11], cell-free (CF) massive MIMO [12]–[15], distributed massive MIMO [16] and large-scale/massive DAS [17], [18], etc. However, it should be noted that the CF massive MIMO considered in [12] is actually a massive distributed MIMO system over a wide area. The phrase ''cell-free'' was used mainly for highlighting the user-centric feature of the mDA systems. The advantages of the so-called mDA system are multi-fold. First, due to the very small distance between AEs and their served users, high spectral-efficiency and/or high energy-efficiency communications can be supported. Second, power-control is only required for quality-of-service (QoS), not for mitigation of near-far problem. Third, handover may become relatively easy, since antenna handover can be used to replace the conventional user handover. Additionally, owing to the distributed inputs and distributed outputs, as well as the distributed signal processing, mDA systems can be flexibly designed to support low-latency communications or/and to provide location-based services.

In recent years, the concept of the so-called mDA has drawn a lot of attention in research, and has been investigated from different perspectives. First of all, to understand the potentials of mDA systems, the system capacity and spectral efficiency have been widely studied, as evidenced, e.g., by [4], [12], [13], [16], [18]–[24] and the reference therein. Among the above-mentioned references, [12], [13], [19] have further compared the spectral-efficiency of mDA systems with that of AP-centric systems. In particular, [13] has shown that CF massive MIMO has the potential to outperform the conventional cellular massive MIMO and small cell networks, by employing the global or local combining techniques in the principle of minimum mean-square error (MMSE). When taking the energy consumption into account, [25], [26] have analyzed the energy-efficiency of mDA systems. Additionally, the authors of [27] have derived an expression providing close approximation of the signal-to-noise ratio (SNR), when global zero-forcing (ZF) detection is employed. In [28], the statistical characteristics of interference in mDA networks have been studied. Second, based on the user-centric setting, the references of [24], [29]–[36] have studied the selection or grouping of AEs, so as to promote the achievable performance of mDA systems. Specifically, the problem of AE grouping can be solved via resource allocation [29], [31], or precoding design [34]–[36]. Furthermore, the joint optimization of resource allocation and precoding design in mDA systems can be implemented [14], [37], [38]. Different from the most of existing works assuming time-division duplex (TDD) in mDA, the frequency-division duplex (FDD)-based CF massive MIMO systems have been considered in [15]. In [39], the message passing algorithm (MPA) has been introduced to implement the ZF beamforming in a large scale cellular network. Besides the above-mentioned topics, the issues of physical layer security [40], [41] and energy harvesting [17], [42] have also been studied in the context of mDA systems.

This article focuses on the uplink multiuser detection (MUD) in the mDA systems, which has so far received relatively less attention in research. For a massive MIMO system with colocated AEs, it is well known that owing to the orthogonality among the different users' channels, the single user detection based on, such as, maximal ratio combining (MRC) is capable of achieving near-optimum performance. However, in a mDA system, although there is a massive number of AEs, due to the distributed nature of AEs and the path-loss of transmitted signals, the actual number of AEs used to distinguish a desired user should be much less than the total number of AEs. Hence, the MRC-based single user detection may be inefficient. On the other hand, the multiuser detection exploiting full-AE cooperation would result in extremely high complexity and the huge resource consumption for information exchange among AEs and CPU. These considerations motivate us to design the efficient MUD schemes, which are able to provide a well-balanced trade-off between complexity and performance. In this context, the MUD based on local signal processing is desirable in mDA systems. However, the feasibility of employing the local signal processing based MUD should first be comprehensively investigated. Furthermore, the MUD should be well designed in order to avoid significant performance loss in comparison with the global one of high-complexity.

Additionally, it is well-known that channel state information (CSI) plays a critical role in modern wireless communications. However, due to the distributed characteristic of AEs, the channel estimation in mDA systems is very different from that in the AE colocated massive MIMO systems. In mDA systems, an AE can only estimate the channels from its nearby users, while the channels from far away users are hard to estimate due to the fact that the received signals from far away users are weak. Hence, when considering all the users as well as all the AEs in a mDA system, it can be inferred that the aggregate channel matrix of the mDA system is a sparse matrix. Thanks to this sparsity, the signal processing methods originating from the compressed sensing (CS) and the MPA may be exploited to design low-complexity MUDs.

With the above-mentioned motivation and consideration, the contributions of this article can be summarized as follows:

- A mDA system is proposed and studied. Based on the theory of stochastic geometry, insightful analysis is carried out, showing that local detection may achieve promising performance at much lower complexity and also at much lower demand on information interchange than global detection. Furthermore, the sparsity

of the aggregate channel matrix in mDA systems is analyzed.

- The orthogonal matching pursuit (OMP) based channel estimation method is proposed. Based on the proposed method, each of the AEs is only exploited to estimate the channels from its nearby users not experiencing deep fading. Hence, multiuser diversity gain can be attained. Furthermore, the factor graph describing the connections between AEs and users can be derived, based on which high-efficiency low-complexity signal detection schemes are designed.
- By exploiting the sparsity of the aggregate channel matrix, MPA-MUD is proposed. When empowered by MPA-MUD, a mDA system is capable of distributing most of the computations among the AEs, while the CPU is only required to maintain the factor graph, and to implements some addition operations.

The rest of the paper is organized as follows. In Section II, the proposed mDA system model is described. Then, based on the theory of stochastic geometry, some insightful analysis for the performance of mDA systems is provided in Section III. In Section IV and Section V, the channel estimation and MUDs are addressed, respectively. In Section VI, some simulation results are provided to illustrate the achievable BER performance of mDA systems. Finally, Section VII summarizes the main conclusions of our research.

II. SYSTEM MODEL

We consider a mDA system, where a massive number of AEs are distributed within an area to be studied. Following the principles of massive MIMO, we assume that the density of AEs is higher than the density of the users supported by the system. Hence, different from the conventional cellular systems, in which the operations are AP-centric, user-centric operations are more effective in this mDA system [43]. In this mDA system, each AE can be viewed as an AP or a remote radio head (RRH), which may serve all the users in the system or just the users around it, depending on the specific transmission/receiving strategies introduced, as illustrated, e.g, in our forthcoming discourses. We assume that all AEs are connected to a CPU via backhaul links, which are assumed to be ideal in terms of reliability and delay. We assume that CPU implements global signal processing, while AEs implement the functions of conveying signals from radio frequency (RF) to baseband or from baseband to RF, as well as local signal processing whenever necessary. Furthermore, each mobile user is assumed to be equipped with a single antenna. As an example, Fig.1 shows such a mDA system, which consists of 13 distributed antennas simultaneously serving 5 users in the same time-frequency resource block. In this article, we model the mDA system based on the stochastic geometry theory [44]. Correspondingly, AEs are distributed according to the homogeneous Poisson point process (PPP) Φ_a with the intensity of λ_a, while users according to the homogeneous PPP Φ_u with the intensity of λ_u, both in the Euclidean plane. Following the above assumption, we have $\lambda_a \gg \lambda_u$.

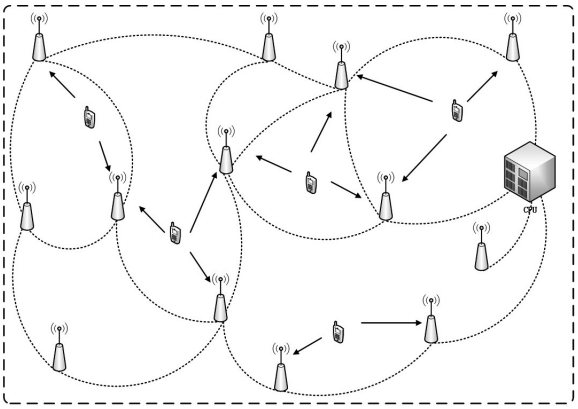


FIGURE 1. An example to illustrate the concept of mDA systems.

In this article, we focus on the uplink synchronous transmission as in [10], [12].[1] Correspondingly, the uplink during each coherence time period is divided into two phases, one of which is used for channel training, and the other one is for data transmission. During the channel training phase, all users send their pilot sequences to the distributed AEs, and based on which each AE estimates the CSI of the users connected to the AE with sufficient signal strength. After the training phase, users then send their datas, which are detected by the CPU/AEs with the aid of the estimated CSI, as detailed in the forthcoming discourses. Let $(\tau + T)$ be the length of a coherence time period, where τ and T are the number of samples used to transmit pilot sequences and data symbols, respectively. Furthermore, we assume that for a reference realization of the mDA system, the number of AEs and users are N and K, respectively.[2] Based on these assumptions and settings, the received signals during the training phase can be expressed as

$$\boldsymbol{Y}_0 = \boldsymbol{\Phi}\boldsymbol{H}^T + \boldsymbol{N}_0 \tag{1}$$

where $\boldsymbol{Y}_0$ is a $(\tau \times N)$ matrix. The n-th column of $\boldsymbol{Y}_0$ gives the observations obtained from the n-th AE over the training period, when the pilots $\boldsymbol{\Phi} = [\boldsymbol{\phi}_1, \boldsymbol{\phi}_2, \ldots, \boldsymbol{\phi}_K]$ are transmitted, where $\boldsymbol{\phi}_k$ is a τ-length pilot sequence of user k. In this article, for the sake of brevity and focusing our attention on the more important issues related to distributed antennas, we assume frequency-nonselective (fast) fading channels. Consequently, the channel matrix $\boldsymbol{H}$ in (1) is a $(N \times K)$ matrix, whose (n, k)-th entry, denoted by $h_{n,k}$, is the channel gain from the k-th user to the n-th AE. Finally, $\boldsymbol{N}_0$ in (1) is the noise matrix of size $(\tau \times N)$, the entries of which are independent and identically distributed (i.i.d.) zero-mean complex Gaussian random variables with a common variance of σ_0^2. Based on the observation matrix of $\boldsymbol{Y}_0$ and the knowledge about the pilot sequences $\boldsymbol{\Phi}$, the CSI of K users may be estimated either jointly by the CPU or separately by each of the AEs.

[1] Strictly speaking, the system is asynchronous due to the dense deployment, where users are connected with their nearby AEs. However, the system can be approximately treated to be synchronous for analysis [10], [12].

[2] Notice: both N and K are random variables, following the PPP models as assumed above.

After the channel training phase, users transmit their data symbols to AEs during the data transmission phase. Correspondingly, the observation matrix produced by AEs can be expressed as

$$\boldsymbol{Y}_1 = \boldsymbol{H}\boldsymbol{X} + \boldsymbol{N}_1 \tag{2}$$

where $\boldsymbol{Y}_1$ is $(N \times T)$-dimensional, $\boldsymbol{H}$ is the same as that in (1) due to the same coherence period is considered, $\boldsymbol{X}$ is a $(K \times T)$ data symbol matrix, whose (k, t)-th entry, i.e, $x_{k,t}$, denotes the data symbol sent by the k-th user on the t-th sample. The power of $x_{k,t}$ is expressed as P_t. Similar to (1), $\boldsymbol{N}_1$ in (2) is a $(N \times T)$ Gaussian noise matrix, and its entries are i.i.d complex Gaussian random variables with zero-mean and a common variance of σ_1^2. Note that, $\boldsymbol{Y}_0$ in (1) and $\boldsymbol{Y}_1$ in (2) are in the form of the transposes of each other, i.e., the columns of $\boldsymbol{Y}_0$ and the rows of $\boldsymbol{Y}_1$ correspond to different AEs, while the rows of $\boldsymbol{Y}_0$ and the columns of $\boldsymbol{Y}_1$ correspond to different signal samples. Furthermore, σ_1^2 can in general be assumed to be equal to σ_0^2 in most cases.

Considering the distribution characteristics of AEs and users, the channel gain between an AE and a user can be modeled by a composite shadowing fading model [45], where shadowing slow fading obeys lognormal distribution, while fast fading follows Rayleigh distribution. Specifically, the channel gain $h_{n,k}$ between the n-th AE and the k-th user can be written as $h_{n,k} = \sqrt{\beta_{n,k}}\alpha_{n,k}$, when taking into account of propagation path-loss, shadowing, and fast fading. Here, $\alpha_{n,k}$ represents the fast-fading factor, obeying the i.i.d. complex Gaussian distribution with zero-mean and a variance of 1, while $\beta_{n,k}$ obeys the lognormal distribution with the probability density function (PDF) given by

$$p_{\beta_{n,k}}(x) = \frac{\xi}{\sqrt{2\pi}\sigma_\beta x} \exp\left[-\frac{(10\log_{10} x - \mu_{n,k})^2}{2\sigma_\beta^2}\right] \tag{3}$$

where $\xi = 10/\ln 10 = 4.3429$, and $\mu_{n,k}$ (dB) and σ_β (dB) are the mean and standard deviation of $10\log_{10}\beta_{n,k}$, respectively. Furthermore, $\mu_{n,k}$ accounts for the propagation path-loss, which is modeled by the double-slope path-loss model [46] formulated by

$$\mu_{n,k}(r) = \begin{cases} 0, & r < 1 \\ -10\log_{10}\left[r^{\nu_1}\left(1 + \frac{r}{g}\right)^{\nu_2}\right] + C, & r \geqslant 1 \end{cases} \tag{4}$$

where r is the distance relative to a reference distance, ν_1 is the basic pass-loss exponent (approximately two), ν_2 is the additional pass-loss exponent, which may range from 2 to 6, g is called the break point of the path-loss curve, and $C = 10\log_{10}(1 + 1/g)^{\nu_2}$ in order to make $\mu_{n,k}(r = 1) = 0$ dB.

III. LOCAL DETECTION VERSUS GLOBAL DETECTION

In the considered mDA systems, AEs and users are supposed to be distributed within a large area. Hence, signal detection may be implemented globally or locally, forming the global detection or local detection. In the context of global detection, the uplink signals collected by all the AEs are forwarded to the CPU, where signal detection is implemented with the aid of the channel knowledge about all the users. In this case, each individual AE contributes to the detection of all the users in a mDA system. By contrast, when local detection is implemented, a subset of users may only be served by their nearby AEs, and correspondingly, their signal detection can be achieved either at the AEs or at the CPU by employing only the CSI between this subset of users and the AEs serving them.

In this section, we provide some insightful analysis for the performance of mDA systems, when both the local detection and global detection are respectively considered. For the sake of simplicity, we ignore the shadowing effect at moment in our analysis, since it is not very significant as shown in [47]. The propagation path-loss is modeled as $r^{-\nu}$, when $r > 1$, and otherwise, the path-loss factor is equal to 1 (0 dB), when $r \leq 1$, i.e., let $\nu_1 = \nu$ and $\nu_2 = 0$ in (4). Hence, given A of the area covered by a mDA system, the average number of AEs and that of users are $A\lambda_a$ and $A\lambda_u$, respectively. Let us start the analysis.

A. SIGNAL-TO-LEAKAGE RATIO IN mDA SYSTEMS

In order to show the feasibility of employing the local detection in mDA systems, for a given user, we investigate the ratio of $\mathbb{E}[P_{\mathrm{r}}(r)/\bar{P}_{\mathrm{r}}(r)]$, where $P_{\mathrm{r}}(r)$ and $\bar{P}_{\mathrm{r}}(r)$ are respectively the power received from the user by the AEs having their distance less than r from the user, and the power received from the user by the AEs having their distances larger than r from the user. When local detection is employed, i.e., when the user is only served by the AEs having their distances less than r from the user, $P_{\mathrm{r}}(r)$ is the useful power for detecting the desired signals. By contrast, $\bar{P}_{\mathrm{r}}(r)$ can be viewed as the interfering power to the other AEs, namely the leakage power of the user. Hence, the ratio of $\mathbb{E}[P_{\mathrm{r}}(r)/\bar{P}_{\mathrm{r}}(r)]$ is defined as the signal-to-leakage ratio (SLR). Obviously, if SLR is high, most of the user's transmit power is utilized for its detection, and only a small fraction of its transmit power is leaked out to interfere the users associated with the other AEs.

Let us assume that the area covered by a mDA system is infinite, and according to the PPP, we can further assume that the user considered, which is referred to as the reference user, is located at the origin [44]. Since the area contributing $P_{\mathrm{r}}(r)$ is disjoint with the area yielding $\bar{P}_{\mathrm{r}}(r)$, according to the definition of PPP, $P_{\mathrm{r}}(r)$ and $\bar{P}_{\mathrm{r}}(r)$ are independent. Then, we have $\mathbb{E}[P_{\mathrm{r}}(r)/\bar{P}_{\mathrm{r}}(r)] = \mathbb{E}[P_{\mathrm{r}}(r)]\mathbb{E}[1/\bar{P}_{\mathrm{r}}(r)]$. According to [44], we can readily show that

$$\mathbb{E}[P_{\mathrm{r}}(r)] = \lambda_a\pi + \frac{2\lambda_a\pi}{\nu - 2}(1 - r^{2-\nu}) \tag{5}$$

when the transmit power of reference user is normalized to one. In order to derive $\mathbb{E}[1/\bar{P}_{\mathrm{r}}(r)]$, let us introduce an auxiliary variable a, which obeys the exponential distribution with unit mean, and is independent of $\bar{P}_{\mathrm{r}}(r)$. Then, we have $\mathbb{E}[1/\bar{P}_{\mathrm{r}}(r)] = \mathbb{E}[a/\bar{P}_{\mathrm{r}}(r)]$. Furthermore, for given $\bar{P}_{\mathrm{r}}(r)$, we have $P(a/\bar{P}_{\mathrm{r}}(r) > s) = P(a > s\bar{P}_{\mathrm{r}}(r)) = \exp(-s\bar{P}_{\mathrm{r}}(r))$, owing to a obeying the exponential distribution with

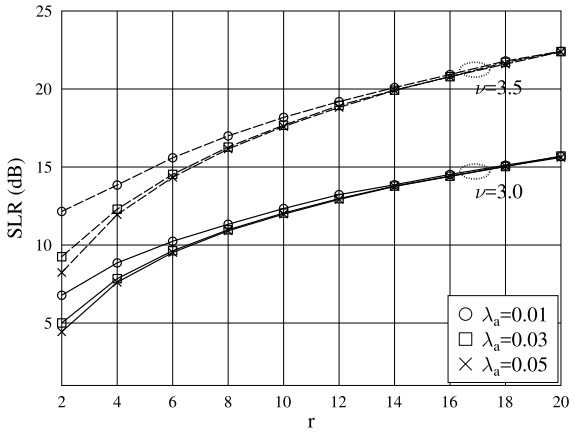


FIGURE 2. SLR versus r when $\nu = 3$ and $\nu = 3.5$.

unity mean. Let the complementary cumulative distribution function (CCDF) of the variable $a/\bar{P}_{\rm r}(r)$ be denoted as $\bar{F}(s)$, i.e., $\bar{F}(s) = P(a/\bar{P}_{\rm r}(r) > s) = \mathbb{E}[\exp(-s\bar{P}_{\rm r}(r))]$, which is the expectation of $\exp(-s\bar{P}_{\rm r}(r))$ with respect to $\bar{P}_{\rm r}(r)$, since $\bar{P}_{\rm r}(r)$ is also a random variable. Then from the above analysis we have

$$\mathbb{E}\left[\frac{1}{\bar{P}_{\rm r}(r)}\right] = \mathbb{E}\left[\frac{a}{\bar{P}_{\rm r}(r)}\right] = -\int_0^\infty s \cdot d\bar{F}(s) \tag{6}$$

Explicitly, $\bar{F}(s) = \mathbb{E}[\exp(-s\bar{P}_{\rm r}(r))]$ is the Laplace transform (LT) of $\bar{P}_{\rm r}(r)$, which is derived in Appendix A. Specifically for $r \geqslant 1$, we have $\bar{F}(s) = \mathcal{L}_{P_{\rm r}(r)}(s)$, and

$$\mathcal{L}_{P_{\rm r}(r)}(s) = \exp\left\{-s^\delta \frac{\lambda_a \pi}{\mathrm{sinc}(\delta)} + \lambda_a \pi r^2 \left[1 - \frac{1}{(sr^{-\nu}+1)} \times \frac{\delta}{(\delta+1)}\, {}_2F_1\left(1, 1; \delta+2; \frac{1}{sr^{-\nu}+1}\right)\right]\right\} \tag{7}$$

where $\delta = 2/\nu$, $\mathrm{sinc}(\delta) = \sin(\pi\delta)/\pi\delta$, and ${}_2F_1(a, b; c; x)$ is the Gauss' Hypergeometric function [48]. Finally, from (5) and (6) we have the SLR expressed as

$$\mathbb{E}\left[\frac{P_{\rm r}(r)}{\bar{P}_{\rm r}(r)}\right] = \left(\lambda_a \pi + \frac{2\lambda_a \pi}{\nu - 2}(1 - r^{2-\nu})\right) \cdot \left(-\int_0^\infty s \cdot d\bar{F}(s)\right) \tag{8}$$

In order to obtain some intuitive results, we apply the Jensen's inequality to obtain $1/\mathbb{E}[\bar{P}_{\rm r}(r)] \leqslant \mathbb{E}[1/\bar{P}_{\rm r}(r)]$. Therefore, we have $\mathbb{E}[P_{\rm r}(r)/\bar{P}_{\rm r}(r)] \geq \mathbb{E}[P_{\rm r}(r)]/\mathbb{E}[\bar{P}_{\rm r}(r)]$. Furthermore, since the total power received by a mDA system under our assumption is $\mathbb{E}[P_{\rm r}] = \lambda_a \pi + 2\lambda_a \pi/(\nu - 2)$ [44], we have $\mathbb{E}[\bar{P}_{\rm r}(r)] = \mathbb{E}[P_{\rm r}] - \mathbb{E}[P_{\rm r}(r)] = 2\lambda_a \pi r^{2-\nu}/(\nu - 2)$. Consequently, the SLR can be approximately be given as

$$\mathbb{E}\left[\frac{P_{\rm r}(r)}{\bar{P}_{\rm r}(r)}\right] \geqslant \frac{\mathbb{E}[P_{\rm r}(r)]}{\mathbb{E}[\bar{P}_{\rm r}(r)]} = \frac{\nu}{2r^{2-\nu}} - 1 \tag{9}$$

Fig.2 shows the SLR versus distance r. From the results as well as (8) and (9), we can deduce that the majority of power from a user can be received by the AEs near the user. For example, assume that $\nu = 3.5$ and $\lambda_a = 0.03$. Then, we have $\mathbb{E}[P_r(r)/\bar{P}_r(r)] \approx 14.5$ dB for $r = 6$ and 19 dB for $r = 12$. This observation in turn implies that the interference on an AE is mainly from the users around the AE, where the interference from the users far away from the AE is insignificant. Therefore, when the value of ν is relatively large, meaning severe path-loss, we can expect that the signal detection based on the local observations of a user is close to that based on the global observations obtained from all the AEs in the system. Additionally, as seen in (9) and Fig.2, the intensity of AEs quantified by λ_a has only slight impact on the SLR. This is the direct result of the homogeneous PPP modeling of AEs.

B. SINGLE-USER SCENARIOS

When there is just one user in the system, we can assume that the user is positioned at the origin without any penalization of generality. Let us assume for simplicity that binary phase-shift keying (BPSK) modulation is employed. Then, the MRC scheme is optimum for detection, which yields the average BER of

$$P_b = \mathbb{E}\left[Q\left(\sqrt{2\gamma_t \sum_{n=1}^{N} \beta_{n,1}|\alpha_{1,n}|^2}\right)\right] \tag{10}$$

where $\gamma_t = P_t/\sigma_1^2$ is the ratio between the signal's transmit power P_t and the noise power σ_1^2 presenting at AEs, or the noise-variance normalized transmit power. Hence, we refer to it as the transmit SNR (TSNR). In (10), $Q(x)$ is the Gaussian Q-function. Upon invoking the definition of $Q(x) = \frac{1}{\pi}\int_0^{\pi/2} \exp\left(-\frac{x^2}{2\sin^2\theta}\right) d\theta$, $x \geqslant 0$ [49], the average BER can be expressed as

$$P_b = \frac{1}{\pi}\int_0^{\pi/2} \mathbb{E}\left[\exp\left(-\frac{\gamma_t \sum_{n=1}^{N} \beta_{n,1}|\alpha_{1,n}|^2}{\sin^2\theta}\right)\right] d\theta \tag{11}$$

Letting $I = \sum_{n=1}^{N} \beta_{n,1}|\alpha_{1,n}|^2$, we can express the expectation in (11) as $\mathbb{E}\left[\exp\left(-\gamma_t I/\sin^2\theta\right)\right]$, which is explicitly the LT of I with $s = \gamma_t/\sin^2\theta$. Then, when assuming that the area covered by a mDA system is infinite, from the stochastic geometry theory [44] we can obtain

$$P_b = \frac{1}{\pi}\int_0^{\pi/2} e^{\Lambda(\theta)} d\theta \tag{12}$$

where $\Lambda(\theta)$ for global and local detection is derived in Appendix A by letting $s = \gamma_t/\sin^2\theta$. Specifically, when the global detection is implemented, $\Lambda(\theta) \overset{\rm def}{=} \Lambda_G(\theta)$, which is given by (32). By contrast, when the local detection is implemented, we have $\Lambda(\theta) \overset{\rm def}{=} \Lambda_L(\theta)$, which is given by (33).

In order to remove the integration seen in (12) and observe clearly the effect of the involved parameters, we may employ the approximation of $Q(x) \approx e^{-\frac{x^2}{2}}/12 + e^{-\frac{2x^2}{3}}/4$ and the Chernoff bound of $Q(x) \leqslant e^{-\frac{x^2}{2}}$ to simplify (12) [49]. Note that, the Chernoff bound can be readily obtained from (11) by letting $\sin^2\theta = 1$. Then, for employing global detection, we have $P_b \approx e^{\Lambda_G}/12 + e^{\frac{4}{3}\Lambda_G}/4 \leqslant e^{\Lambda_G}$, where Λ_G is

obtained from $\Lambda_G(\theta)$ by letting $\sin^2\theta = 1$, resulting in

$$\Lambda_G = \frac{\lambda_a \pi}{\gamma_t + 1} - \frac{\lambda_a \pi \delta}{(\delta+1)(\gamma_t+1)}\, {}_2F_1\left(1, 1; \delta+2; \frac{1}{\gamma_t+1}\right) - \gamma_t^{\delta} \frac{\lambda_a \pi}{\mathrm{sinc}(\delta)} \quad (13)$$

Similarly, when employing local detection, we have $P_b \approx e^{\Lambda_L}/12 + e^{\frac{4}{3}\Lambda_L}/4 \leqslant e^{\Lambda_L}$, where Λ_L is obtained from $\Lambda_L(\theta)$, given by

$$\Lambda_L = \frac{\lambda_a \pi}{\gamma_t + 1} - \frac{\lambda_a \pi \delta}{(\delta+1)(\gamma_t+1)}\, {}_2F_1\left(1, 1; \delta+2; \frac{1}{\gamma_t+1}\right) + \frac{\lambda_a \pi \delta R^2}{(\delta+1)(\gamma_t R^{-\nu}+1)}\, {}_2F_1\left(1, 1; \delta+2; \frac{1}{\gamma_t R^{-\nu}+1}\right) - \lambda_a \pi R^2 \quad (14)$$

Now let us have a close look at Λ_G and Λ_L, which determine the single-user BER bound of the mDA systems employing respectively the global and local detection. In the formulas of (13) and (14), γ_t is the TSNR, as above-mentioned, which in practice is usually very big, typically > 100 dB. Therefore, the first two terms in Λ_G and Λ_L are very close to 0. Consequently, we can have a very close approximation of $P_b \propto (e^{-\gamma_t^{\delta}})^{\lambda_a \pi/\mathrm{sinc}\delta}$ for the mDA systems employing the global detection.

By contrast, when the local detection is considered, for a given R, there is an error floor of $P_b \propto e^{-\lambda_a \pi R^2}$, when the TSNR is sufficiently high to make all the other γ_t-related terms approximately zero. Note that, in $P_b \propto e^{-\lambda_a \pi R^2}$, $\lambda_a \pi R^2$ is actually the average number of AEs located in a circular area with a radius of r, where the reference user is at the centre of this area. Hence, this error floor is caused by the possibility that there are no AEs in this circular area.

C. MULTI-USER SCENARIOS

In the multiuser communications scenarios, without any loss of generality, we assume that the first user is the reference user, and the data transmitted within the first time-slot by the reference user is detected. Following the studies in Section III-B, here we consider only the local detection of the reference user based on the observations received from the set of AEs expressed as $\mathcal{I}_1$. Similar to the single-user case, an AE serving the reference user should satisfy the requirement that its distance from the reference user is less than a threshold denoted by r. Correspondingly, the observations collected by the AEs in $\mathcal{I}_1$ can be expressed as

$$\boldsymbol{y}^{[1]} = \boldsymbol{H}^{[1]} \boldsymbol{x}_1 + \boldsymbol{n}^{[1]} \quad (15)$$

where $\boldsymbol{x}_1$ is the first column of $\boldsymbol{X}$ seen in (2), and $\boldsymbol{H}^{[1]}$ is a $(|\mathcal{I}_1| \times K)$ matrix constructed by the channel gains from all the K users to the AEs serving the reference user. Note that since the symbols transmitted by users are assumed to be independent and detection is carried out slot-by-slot, without any loss of generality, we can only consider the symbols transmitted in the first slot, i.e., $\boldsymbol{x}_1$, in the analysis. Furthermore, when local detection is considered, the number of users K may be more than the number of AEs in $|\mathcal{I}_1|$. Finally, $\boldsymbol{n}^{[1]}$ in (15) is a $(|\mathcal{I}_1| \times 1)$ complex Gaussian vector with zero mean and a covariance matrix $\sigma_1^2 \boldsymbol{I}_{|\mathcal{I}_1|}$.

When the correlation based detection is employed, the decision variable for detecting the reference user is

$$\begin{aligned} z_1 &= \left(\boldsymbol{h}_1^{[1]}\right)^H \boldsymbol{y}^{[1]} \\ &= \left(\boldsymbol{h}_1^{[1]}\right)^H \boldsymbol{h}_1^{[1]} x_{1,1} + \sum_{k=2}^{K} \left(\boldsymbol{h}_1^{[1]}\right)^H \boldsymbol{h}_k^{[1]} x_{k,1} + \left(\boldsymbol{h}_1^{[1]}\right)^H \boldsymbol{n}^{[1]} \end{aligned} \quad (16)$$

where $\boldsymbol{h}_k^{[1]}$ is the k-th column of $\boldsymbol{H}^{[1]}$. From (16) we can conceive that due to the propagation path-loss and our assumption of $\lambda_a > \lambda_u$, there is a high probability that $||\boldsymbol{h}_1^{[1]}||_2 > ||\boldsymbol{h}_k^{[1]}||_2$ for any $k \neq 1$, which is more declared, when λ_a increases. This implies that when the reference user is detected based on the observations sampled from its near-field AEs, the reference user is in general the strongest user. Therefore, even the single-user correlation detector is feasible, and does not experience near-far problem. Additionally, we note that the local detection tends to the global detection, when $r \to \infty$.

Let us assume that the area covered by a mDA system is infinite. This assumption makes that any AE in $\mathcal{I}_1$ conflicts in average the same amount of interference power, expressed as

$$\sigma_{IN}^2 = \sigma_1^2 + \sum_{k=2}^{K\to\infty} \mathbb{E}\left[|h_{m,k}|^2\right] = \sigma_1^2 + \sum_{k=2}^{K\to\infty} \mathbb{E}\left[\beta_{m,k}\right] \quad (17)$$

Then, when users are assumed to be uniformly distributed with the intensity λ_u, it can be shown that $\sum_{k=2}^{K\to\infty} \mathbb{E}\left[\beta_{m,k}\right] = P_t\,(\lambda_u \pi + 2\lambda_u \pi/(\nu-2))$. Consequently, when adopting the Gaussian approximation on (16), the BER of the mDA systems employing BPSK baseband modulation is given by

$$P_b = \mathbb{E}\left[Q\left(\sqrt{2\gamma_{It} \sum_{n=1}^{N} \beta_{n,1} |\alpha_{1,n}|^2}\right)\right] \quad (18)$$

where $\gamma_{It} = P_t/\sigma_{IN}^2$, which is the transmit power normalized by the interference-plus-noise power at AE, referred to as the transmit signal to interference-plus-noise ratio (TSINR). Explicitly, (18) has the same characteristics as (10). Hence, the BER of (18) can be evaluated from the formula of (12) and the other related formulas by letting $s = P_t/(\sigma_{IN}^2 \sin^2\theta)$, or from the approximation of the Q-function by replacing γ_t by γ_{It}.

D. SPARSITY OF H

From Section III-A we can know that owing to the propagation path-loss, most of the power received from a user in a mDA system is contributed by the AEs near the user. This implies that some entries in the channel matrix $\boldsymbol{H}$ are dominant, while the others are insignificant. Therefore, $\boldsymbol{H}$ can be regarded as a quasi-sparse matrix, as shown below.

Specifically for the k-th user in a mDA system, the channel gains from it to all the AEs in the system is a vector, which is the k-th column of $\boldsymbol{H}$. Let us denote the k-th user's channel vector as $\boldsymbol{h}_k$. Assume that the area covered by the mDA system is large enough. Then, the PDF of the distance between the k-th user and its i-th nearest AE can be derived from the Poisson distribution by assuming that there are at least $(i-1)$ AEs in the area of πr^2. For given i, this PDF can be expressed as [44]

$$f_i(r) = \frac{2}{\Gamma(i)}(\lambda_a \pi)^i r^{2i-1} e^{-\lambda_a \pi r^2}, \ r \geq 0 \tag{19}$$

With the aid of this PDF and applying the independence assumption between small-scale fading and large-scale pathloss, we have

$$\begin{aligned}
&\mathbb{E}[|h_{n(i),k}|^2] \\
&= \int_0^1 f_i(r)dr + \int_1^\infty r^{-\nu} f_i(r)dr \\
&= \frac{1}{\Gamma(i)}\left[\gamma(i, \lambda_a\pi) + (\lambda_a\pi)^{\frac{\nu}{2}}\Gamma\left(i - \frac{\nu}{2}, \lambda_a\pi\right)\right]
\end{aligned} \tag{20}$$

where $\Gamma(\cdot)$, $\Gamma(\cdot,\cdot)$ and $\gamma(\cdot,\cdot)$ are the gamma function, the upper incomplete gamma function and the lower incomplete gamma function, respectively. Hence, when considering the first J nearest AEs of user k, we have

$$\begin{aligned}
&\mathbb{E}\left[\sum_{i=1}^{J} |h_{n(i),k}|^2\right] \\
&= \sum_{i=1}^{J} \mathbb{E}[|h_{n(i),k}|^2] \\
&= \sum_{i=1}^{J} \frac{1}{\Gamma(i)}\left[\gamma(i, \lambda_a\pi) + (\lambda_a\pi)^{\frac{\nu}{2}}\Gamma\left(i - \frac{\nu}{2}, \lambda_a\pi\right)\right]
\end{aligned} \tag{21}$$

On the other side, we have $\boldsymbol{h}_k^H \boldsymbol{h}_k = \sum_{n=1}^{N} \beta_{n,k}|\alpha_{n,k}|^2$, where N is the total number of AEs in the mDA system. Then, as $N \to \infty$, or equivalently, as the area covered by a mDA system reaches infinite, from (5) and letting $r \to \infty$, we have

$$\mathbb{E}[\boldsymbol{h}_k^H \boldsymbol{h}_k] = \mathbb{E}[||\boldsymbol{h}_k||_2^2] = \lambda_a\pi + \frac{2\lambda_a\pi}{\nu - 2} \tag{22}$$

When comparing (21) and (22), we can show that the channel power contributed by the first J nearest AEs of a user are dominant, provided that $J \geq 3$. Specifically, when assuming $\lambda_a = 0.05$, $\nu = 3.5$ and $J = 3$, from (21) we obtain $\sum_{i=1}^{J=3} \mathbb{E}[\beta_{n(i),k}|\alpha_{n(i),k}|^2] = 0.3369$. By contrast, the total channel power collected from all the AEs is $\mathbb{E}[||\boldsymbol{h}_k||_2^2] = 0.3665$ according to (22). Therefore, using 3 nearest AEs of a user can collect about $0.3369/0.3665 = 92\%$ of the total power available, only about 8% of the total power is contributed by all the AEs other than the three nearest ones, which may hence be ignored. In terms of $\boldsymbol{H}$, the above observation implies that each column of $\boldsymbol{H}$ has only about three dominant elements, while all the other elements can be treated as noise. Hence in mDA systems, $\boldsymbol{H}$ is a sparse matrix, and therefore, sparse signal processing methods [50] may be employed for signal detection in mDA systems, in order to take the advantages of the low-complexity sparse signal processing algorithms. In the following sections, CS-based channel estimation and MPA-MUD are introduced to the mDA systems.

IV. CHANNEL ESTIMATION

In MIMO systems, classical channel estimation methods are typically the maximum-likelihood (ML) estimation and linear minimum mean-square error (LMMSE) estimation, which can be formulated as $\hat{\boldsymbol{H}} = \boldsymbol{W}\boldsymbol{Y}_0$ [51], where $\boldsymbol{W} = (\boldsymbol{\Phi}^H\boldsymbol{\Phi})^{-1}\boldsymbol{\Phi}^H$ in ML estimation, and $\boldsymbol{W} = (\boldsymbol{\Phi}^H\boldsymbol{\Phi} + \sigma_0^2\boldsymbol{I}_K)^{-1}\boldsymbol{\Phi}^H$ in LMMSE estimation. In the above equations, $\boldsymbol{Y}_0$, $\boldsymbol{\Phi}$ and σ_0^2 are all defined in association with (1). As mDA systems belong to the MIMO family, straightforwardly, the above two methods can be adopted for channel estimation in mDA systems. These channel estimation schemes may be implemented jointly at the CPU, but resulting in extreme complexity and extreme information exchange between AEs and CPU. Hence, promising approaches should be that each AE estimates its related channels locally, and informs CPU the estimated channels, only if centralized signal detection is implemented. Furthermore, owing to the sparsity of the channel matrix $\boldsymbol{H}$, for each of the AEs, only a very small number of channels from nearby users need to be estimated.

To describe the local channel estimation, let us consider the n-th AE, its received signal during the channel estimation phase can be expressed as

$$\boldsymbol{y}_{0,n} = \boldsymbol{\Phi}\boldsymbol{h}_n^{\mathrm{c}} + \boldsymbol{n}_{0,n}, \quad n = 1, 2, \ldots, N \tag{23}$$

where $\boldsymbol{y}_{0,n}$ and $\boldsymbol{n}_{0,n}$ are respectively the n-th columns of $\boldsymbol{Y}_0$ and $\boldsymbol{N}_0$ given in (1), $\boldsymbol{h}_n^{\mathrm{c}}$ is the transpose of the n-th row in the channel matrix $\boldsymbol{H}$, i.e., the n-th column of $\boldsymbol{H}^T$, and $\boldsymbol{\Phi}$ is the pilot sequence matrix defined in (1). Due to the sparsity of $\boldsymbol{H}$, $\boldsymbol{h}_n^{\mathrm{c}}$ is a sparse vector. Hence, the channel estimation based on (23) can be viewed as the problem of sparse signal recovery, which has been widely studied in the field of CS [52], [53]. In this article, we employ the orthogonal matching pursuit (OMP) method to estimate $\boldsymbol{h}_n^{\mathrm{c}}$. More precisely, the tasks of our OMP algorithm include identifying and estimating a fraction of the entries in $\boldsymbol{h}_n^{\mathrm{c}}$, which are dominant in terms of the power of $\|\boldsymbol{h}_n^{\mathrm{c}}\|_2^2$. In other words, with the aid of the OMP algorithm, each of the AEs in the mDA system can adaptively select its nearby users without experiencing deep fading and simultaneously, the channels of the selected users are estimated. Therefore, both user association [2] and channel estimation are accomplished simultaneously. In detail, the OMP-based channel estimation algorithm is described as Algorithm 1. Note that, while Step 4 of Algorithm 1 considers the ML-based estimation, it is worth mentioning that the LMMSE-based estimation can be similarly applied.

Note furthermore that, in the above algorithm, the ending condition is that the number of iterations reaches a preset number S, which can be set according to the required performance and the user intensity. In general, better required performance and higher user intensity demand relatively

Algorithm 1 Orthogonal Matching Pursuit Based Channel Estimation

Input: $\boldsymbol{\Phi}, \boldsymbol{y}_{0,n}$, and preset the number of iterations S.

Output: $\hat{\boldsymbol{h}}_n^{\mathrm{c}}, \Omega$.

1: **Initialization:** $\boldsymbol{y}_{\mathrm{res}}^{(0)} = \boldsymbol{y}_{0,n}$, $\Omega^{(0)} = \emptyset$, $\boldsymbol{\Phi}_{\Omega}^{(0)} = \emptyset$, **and** $s = 1$.

2: **while** $s \leqslant S$ **do**

3: **Identification:** $[k^{(s)}, \boldsymbol{\phi}_{k^{(s)}}] = \arg\max_{k,\boldsymbol{\phi}_k} |\boldsymbol{y}_{\mathrm{res}}^{(s-1)H} \boldsymbol{\phi}_k|$, $\Omega^{(s)} = \Omega^{(s-1)} \cup \{k^{(s)}\}$, **and** $\boldsymbol{\Phi}_{\Omega}^{(s)} = \boldsymbol{\Phi}_{\Omega}^{(s-1)} \cup \{\boldsymbol{\phi}_{k^{(s)}}\}$

4: **Estimation:** $\hat{\boldsymbol{h}}_n^{\mathrm{c}(s)} = (\boldsymbol{\Phi}_{\Omega}^{(s)H} \boldsymbol{\Phi}_{\Omega}^{(s)})^{-1} \boldsymbol{\Phi}_{\Omega}^{(s)H} \boldsymbol{y}_{0,n}$

5: **Update:** $\boldsymbol{y}_{\mathrm{res}}^{(s)} = \boldsymbol{y}_{0,n} - \boldsymbol{\Phi}_{\Omega}^{(s)} \hat{\boldsymbol{h}}_n^{\mathrm{c}(s)}$, **and** $s \leftarrow s + 1$

6: **end while**

7: **Let** $\hat{\boldsymbol{h}}_n^{\mathrm{c}} = \boldsymbol{0}_{K\times 1}$. **According to the indices set** $\Omega^{(S)}$, **let the partial entries of** $\hat{\boldsymbol{h}}_n^{\mathrm{c}}$, **whose indices are given in** $\Omega^{(S)}$, **be assigned the values of the corresponding entries of** $\hat{\boldsymbol{h}}_n^{\mathrm{c}(S)}$.

8: **return** $\hat{\boldsymbol{h}}_n^{\mathrm{c}}, \Omega = \Omega^{(S)}$

large value of S, which also results in higher complexity for implementation. However, when S reaches a sufficient value, further increasing it can only result in marginal performance improvement.

Additionally, other termination conditions may be implemented. Below are some alternatives when assuming a preset threshold T_h.

- *Stops if* $\|\boldsymbol{y}_{\mathrm{res}}^{(s)}\|_2 < T_h$: In Algorithm 1, $\boldsymbol{y}_{\mathrm{res}}^{(s)}$ is contributed by the channels not yet estimated and the noise. When $\|\boldsymbol{y}_{\mathrm{res}}^{(s)}\|_2$ is very small, it is likely that the dominant channels have been estimated, and the remaining un-estimated channels belong to the users far away from the AE. Furthermore, if $\|\boldsymbol{y}_{\mathrm{res}}^{(s)}\|_2$ is identified to be less than a preset threshold at the stage of $s = 0$, most probably, there are no nearby users around the AE. In this case, this AE may be set to the idle mode in order to save energy.
- *Stops if* $\|\boldsymbol{\Phi}^H \boldsymbol{y}_{\mathrm{res}}^{(s)}\|_\infty < T_h$: In this case, the resolvable energy along any pilot $\boldsymbol{\phi}_k$ is very small, inferring that the remaining users are all far away from the AE. Similarly, if $\|\boldsymbol{\Phi}^H \boldsymbol{y}_{\mathrm{res}}^{(s)}\|_\infty$ is very small at Stage $s = 0$, all users in the system are likely far away from the AE concerned, and hence this AE can be set to the idle mode.
- *Stops if* $\|\boldsymbol{y}_{\mathrm{res}}^{(s)}\|_2 - \|\boldsymbol{y}_{\mathrm{res}}^{(s-1)}\|_2 < T_h$: When this condition is met, most probably the dominant channels have all been estimated. Note that, this condition is useful in the cases of high background noise and/or many far away users, which make $\|\boldsymbol{y}_{\mathrm{res}}^{(s)}\|_2$ of an AE have a relatively large value, even when there are no dominant users.

Note that, from the above termination conditions we can know that the sparsity parameters, which are usually required in CS and hence needed to be estimated first, are no longer required in our proposed OMP-based algorithm.

Additionally, it is worthy of noting that as in the considered mDA systems, the intensity of AEs is higher than the intensity of users, the number of nearby users of an AE is usually very low. This implies that each AE usually only needs to estimate a very small number of users, typically, one or two. Therefore, Algorithm 1 has very low complexity. Specifically, the complexity of the proposed OMP-based channel estimation has the complexity of $\mathcal{O}(S^2(S+1)^2/4)$, instead of $\mathcal{O}(K^3)$, which is the complexity of the ML or LMMSE estimator. According to the above analysis, we have $S \ll K$, which is practically the fact in general. Furthermore, according to Algorithm 1, information interchange between an AE and the CPU or among AEs is not required.

Besides channel estimation, information provided by Algorithm 1 can also be used for other purposes. For example, a factor graph [54] can be built from the results obtained by Algorithm 1, in order to run the MPA-MUD. In detail, a factor graph can be formed as follows. Let users (more exactly data symbols sent by them) be the variable nodes, and AEs be the check nodes of the factor graph. When the channel between a user and an AE is positively estimated by Algorithm 1, it infers that the user can be served by the AE with relatively good quality. Then, the corresponding variable node and check node in the factor graph is connected using an edge. After considering all the estimated channels, the factor graph for running the MPA-MUD can be formed, which can be maintained at the CPU and updated following each channel estimation process. In section V, we will address the MPA-MUD in detail. Additionally, we note that Algorithm 1 allows an AE to select the users having relatively large channel gains to associate with it. Hence, the users whose signals to an AE experience deep fading will not be connected to the AE, i.e., Algorithm 1 has the capability to attain multiuser diversity.

Discussion - In the OMP-based channel estimation, the matrix $\boldsymbol{\Phi}$ has significant impact on the achievable performance. Ideally, if channel's coherence time $\tau + T$ is long enough to satisfy $\tau \geqslant K$, discrete Fourier transform (DFT) matrix can be used to construct $\boldsymbol{\Phi}$. In this case, orthogonal pilot sequences are assigned to all the K users in the system. However, in the practical mDA systems expected to support a big number of users, it may be the case that $\tau < K$. In these mDA systems, correlated pilot sequences or pilots reuse, have to be implemented. Specifically, when the CS based on OMP algorithm is employed, the columns of $\boldsymbol{\Phi}$ should satisfy the mutual incoherence property (MIP) [55], in order to ideally recover the sparse signals. Hence, when correlated pilot sequences are used, the correlation between any two columns of $\boldsymbol{\Phi}$ should be as small as possible. According to references, e.g., [56], when given τ and K, the design of $\boldsymbol{\Phi}$ can be viewed as a Grassmannian line packing problem, which can be solved by constructing Grassmannian Frames [57]. Furthermore, in such a mDA system, it can be understood that interference is mainly imposed by the geographically adjacent users. Hence, pilots reuse is possible provided that two users assigned the same pilot or two pilots with high

correlation are separated by sufficient distance. Thanks to the GPS and distributed AEs, users' positions in mDA systems can be readily obtained, which can be exploited for pilots allocation [12], [38], [58].

Furthermore, with the knowledge of users' positions, the size of the candidate sets used at the identification stage of the OMP algorithm can be reduced. This can be achieved by removing the users far away from an AE from its candidate set. Furthermore, as the result that the intensity of AEs is higher than the intensity of users, it can also be expected that the candidate set of an AE has small size. Consequently, the runtime of the OMP algorithm shown in Algorithm 1 can be significantly reduced, owing to the involvement of the small-size candidate sets.

V. SIGNAL DETECTION

Signal detection in mDA systems can be carried out either globally at the CPU or locally at the AEs. For the sake of comparison, in this section, we first briefly introduce the MMSE assisted global multiuser detection (MMSE-GMUD) and MMSE based local multiuser detection (MMSE-LMUD). Then, the MPA based local detection is addressed.

A. MMSE BASED GLOBAL MULTIUSER DETECTION

For the MMSE-GMUD, upon invoking the estimated channel matrix $\hat{\boldsymbol{H}}$, the decision statistics for the transmitted symbols during the t-th time slot can be expressed as $\boldsymbol{z}_t = \boldsymbol{W}^H \boldsymbol{y}_t$, where $\boldsymbol{W} = \hat{\boldsymbol{H}}(\hat{\boldsymbol{H}}^H \hat{\boldsymbol{H}} + \sigma_1^2 \boldsymbol{I}_K)^{-1}$, and $\boldsymbol{y}_t$ is the t-th column of $\boldsymbol{Y}_1$.

Since it was assumed that the intensity of AEs is higher than that of users, we can predict that, if the channel matrix can be estimated with high accuracy, the achievable BER performance of MMSE-GMUD should be close to that of the optimal GMUD, such as the ML-based GMUD (ML-GMUD). However, due to the propagation path-loss, it can be expected that the channel matrix entries corresponding to the AEs receiving poor signals may be estimated inaccurately. Therefore, for a given user, the effective number of AEs used for MUI suppression should be much less than N of the total number of AEs in a mDA system, which hence results in performance loss. Furthermore, GMUD has to be implemented at the CPU, which has high complexity, and is also resource greedy for information interchange between CPU and AEs, especially when large mDA systems are considered. Next, let us consider the MMSE-LMUD.

B. MMSE BASED LOCAL MULTIUSER DETECTION

Without any loss of generality, let the first user be the reference user, whose data symbol transmitted during the first time slot is detected. We assume that each user in the mDA system is served by its nearby AEs with their distances less than R from the user.[3] Equivalently, each AE serves all its nearby users with their distances less than r from the AE. It should be noted that, there is a very low probability that a user might not be served by any AE. In this case, the user's quality of service (QoS) requirement may not be met, as discussed in [30]. However, we should also note that the probability of the above-mentioned event can be controlled to be very small. This is because first in the mDA system, it is assumed that the density of AEs is higher than that of users and second, the distance R can be set to a value so that the outage probability of a user without serving AE can be below a pre-set small value. As shown in Section III-C, the signals collected by the AEs serving the reference user can be expressed as (15). However, we should note that the whole $\boldsymbol{H}^{[1]}$ may not be known to any of the AEs serving the reference user. Instead, each of the AEs serving the reference user may only know a few of the entries in $\boldsymbol{H}^{[1]}$. This is because according to Section IV, each of the AEs can only estimate the channels of its nearby users, and the channels from the users far away from the AE are usually hard to estimate due to the propagation path-loss resulted low SNR. Note furthermore that $\mathcal{I}_1$ has been defined in the context of (15) as the set of AEs serving the reference user. Now, let $\mathcal{U}_{1,i}$ be the set of users served by the i-th AE in the set of $\mathcal{I}_1$. Then, it can be shown that the received signal of the i-th AE can expressed as

$$y_i^{[1]} = \sum_{u_i \in \mathcal{U}_{1,i}} h_{i,u_i} x_{u_i} + \sum_{k \notin \mathcal{U}_{1,i}} h_{i,k} x_k + n_i \tag{24}$$

where $y_i^{[1]}$ is the i-th entry of $\boldsymbol{y}^{[1]}$ of (15). In (24), h_{i,u_i} for any $u_i \in \mathcal{U}_{1,i}$ is known to the i-th AE, as the result that the i-th AE serves the users in $\mathcal{U}_{1,i}$. By contrast, the i-th AE has to treat the signals received from the users not in $\mathcal{U}_{1,i}$ as interference, as it does not know the channels from these users. With these consideration, (15) can be rewritten as

$$\boldsymbol{y}^{[1]} = \boldsymbol{H}_{\mathcal{U}}^{[1]} \boldsymbol{x}_{1,\mathcal{U}} + \boldsymbol{J}^{[1]} + \boldsymbol{n}^{[1]} \tag{25}$$

where $\boldsymbol{x}_{1,\mathcal{U}}$ holds the data symbols transmitted by the users in the set of $\mathcal{U} = \bigcup_{i \in \mathcal{I}_1} \mathcal{U}_{1,i}$, $\boldsymbol{H}_{\mathcal{U}}^{(1)}$ is the channel matrix constructed by h_{i,u_i} for $i \in \mathcal{I}_1$ and $u_i \in \mathcal{U}_{1,i}$, and $\boldsymbol{J}^{[1]}$ is the interference vector having the components as shown by the second item at the right-hand side of (24).

For example, as shown in Fig.3, following the above discussion, we have $\mathcal{I}_1 = \{\text{AE}_1, \text{AE}_2, \text{AE}_3\}$, $\mathcal{U}_{1,1} = \{\text{User}_1, \text{User}_2, \text{User}_3, \text{User}_4\}$, $\mathcal{U}_{1,2} = \{\text{User}_1, \text{User}_2\}$, and $\mathcal{U}_{1,3} = \{\text{User}_1, \text{User}_4, \text{User}_5\}$. Hence, we have

$$\boldsymbol{H}_{\mathcal{U}}^{(1)} = \begin{pmatrix} h_{1,1} & h_{1,2} & h_{1,3} & h_{1,4} & 0 \\ h_{2,1} & h_{2,2} & 0 & 0 & 0 \\ h_{3,1} & 0 & 0 & h_{3,4} & h_{3,5} \end{pmatrix} \tag{26}$$

where $h_{n,k}$ is the channel gain from User_k to AE_n. Note that, the matrix $\boldsymbol{H}_{\mathcal{U}}^{(1)}$ has the property that its first column is always full, as this column corresponds to the reference user.

Let $\hat{h}_{i,u_i}$ be the estimate to h_{i,u_i} for all $i \in \mathcal{I}_1$ and $u_i \in \mathcal{U}_{1,i}$. Then, we have the estimate $\hat{\boldsymbol{H}}_{\mathcal{U}}^{[1]}$ to $\boldsymbol{H}_{\mathcal{U}}^{[1]}$. Consequently, when the MMSE-LMUD is employed,

[3]Note that, with the aid of Algorithm 1, the AEs serving a user may be selected according to the channel estimation results, as discussed in Section IV. In this case, however, the cluster of AEs serving a specific user are required to be updated after every round of channel estimation, hence resulting in the increase of the complexity for user scheduling.

the decision variable for detecting the reference user can be expressed as

$$z^{[1]} = \hat{\boldsymbol{h}}_1^{[1]H} \left(\hat{\boldsymbol{H}}_{\mathcal{U}}^{[1]} \hat{\boldsymbol{H}}_{\mathcal{U}}^{[1]H} + \sigma_{IN}^2 \boldsymbol{I}_{|\mathcal{I}_1|} \right)^{-1} \boldsymbol{y}^{[1]} \tag{27}$$

where $\sigma_{IN}^2 = \sigma_1^2 + 2\lambda_u \pi R^{2-\nu}/(\nu - 2)$, obtained by assuming that the area covered by the mDA system is large enough to use the estimation of (5).

The MMSE-LMUD can be either implemented at the CPU or implemented at a specific AE serving the reference user. When it is implemented at the CPU, all AEs in the system need to forward their observations and estimated channels to the CPU, which may consume a lot of backhaul resources. By contrast, when the MMSE-LMUD is implemented at the AE leyer, the AEs serving the reference user can form a cluster, and one of the AEs is selected as the cluster head. Then, all the AEs in a cluster forward their observations and estimated channels to the cluster header, where information of the reference user is detected. In this way, information interchange only occurs among the adjacent AEs. Hence, the complexity for MMSE-LMUD and the information interchange among AEs is much less than that of the MMSE-GMUD. Additionally, we note that, since in our mDA system the number of AEs is much larger than the number of users, one cluster head (AE) in average only needs to serve one user.

C. MPA BASED MUD

As discussed previously in Section III-D, the channel matrix in mDA systems is in general sparse. Hence, the input-output relationship of a mDA system can be described by a sparse factor graph that is similar as that in the low-density parity-check (LDPC) codes [59]. Correspondingly, the relatively low-complexity MPA used in LDPC decoding can be employed for the signal detection in mDA systems in order to achieve near-optimum performance. However, we should realize that, due to the dynamics of users and the random distributions of users and AEs, the factor graphs for mDA systems appear in random fashion. In some cases when the factor graphs are well connected as in LDPC decoding [54], signal detection can be operated in an iterative way, which can usually achieve near-optimum performance. In some other cases when there exist short cycles, degraded performance may be attained. Furthermore, in some cases, the factor graph of a mDA system may even be divided into several component graphs, which are not connected at all. This is the result that some sets of users do not share any AEs with the other sets of users. In this case, information about the users in one component graph cannot be propagated to another component graph. Certainly, the users belonging to different component graphs also do not interfere with each other.

As an example, the factor graph representing a realization of the mDA system shown in Fig.3 can be pictured in Fig.4. As there are no users near the 6-th AE, it can be switched off and be removed from the graph.

Below we detail the MPA-MUD. For the sake of brevity, BPSK is assumed, which can be straightforwardly extended

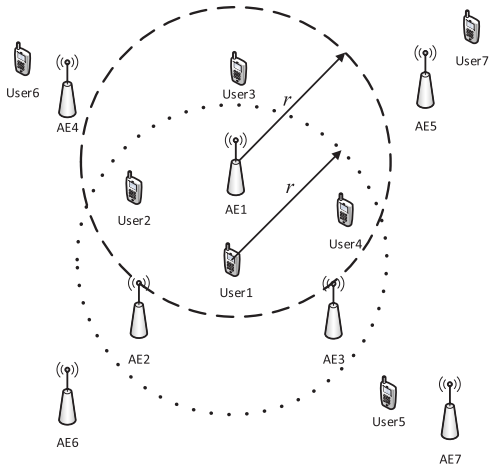


FIGURE 3. An example of the mDA system for employing MMSE-LMUD.

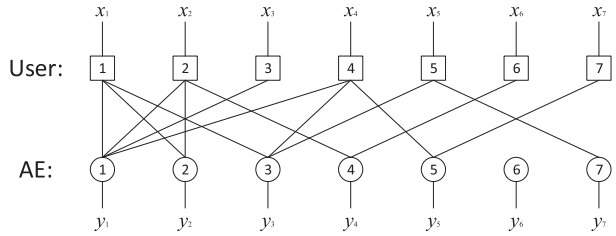


FIGURE 4. An example to illustrate the factor graph representation of the input-output relationship of a mDA system, where y_i is the observation obtained by the i-th AE, and x_j is the data symbol sent by user j.

to the higher order modulation schemes, such as, general quadrature amplitude modulation (QAM) schemes. In order to describe the MPA-MUD, let us define $\mathcal{I}_k$, $k = 1, \ldots, K$, and $\mathcal{U}_n$, $n = 1, \ldots, N$, as the set of AEs serving the k-th user and the set of users served by the n-th AE, respectively. For a given time slot t, the transmitted symbols $x_{k,t}$, $k = 1, \ldots, K$, and corresponding observations $y_{n,t}$, $n = 1, \ldots, N$, are denoted as the variable nodes (VNs) and function nodes (FNs), respectively. Since we assume that $x_{k,t}$ are independent, the MPA-MUD can be implemented slot-by-slot, and hence the subscript t can be ignored without any confusion, i.e., expressing $x_{k,t}$ as x_k, and $y_{n,t}$ as y_n. In this case, $\mathcal{I}_k$ in fact denotes a set of observations of $\{y_n\}$ that are connected to x_k. Similarly, $\mathcal{U}_n$ is a set containing the $\{x_k\}$ that are connected to y_n. Furthermore, let $d_n^f = |\mathcal{U}_n|$ and $d_k^v = |\mathcal{I}_k|$ be the degrees of the n-th FN and k-th VN, respectively.

Based on the above definitions, the MPA-MUD is operated on the factor graph as defined, and messages are updated and exchanged iteratively between the VNs and FNs via the connected edges. Specifically, at the i-th iteration, the message updated by the k-th VN, and to be sent to the n-th FN can be computed by the formula [60]

$$\ell_{n \leftarrow k}^{(i)} = \log \frac{P_{\text{ext},n}(x_k = +1)}{P_{\text{ext},n}(x_k = -1)} = \sum_{m \in \mathcal{I}_k \setminus n} \ell_{m \to k}^{(i-1)} \tag{28}$$

where $P_{\text{ext},n}(x_k) = \eta_{n,k} \exp(x_k \ell_{m \to k}^{(i-1)}/2)$, and $\eta_{n,k}$ is used to make $P_{\text{ext},n}(x_k = +1) + P_{\text{ext},n}(x_k = -1) = 1$. Similarly, the

message updated by the n-th FN and to be sent to the k-th VN can be expressed as [60]

$$\ell_{n\to k}^{(i)} = \log \frac{\sum_{\substack{\boldsymbol{x}^{[n]}\in\mathbb{X}^{d_n^f} \\ x_k=+1}} \exp\left(\sum_{l\in\mathcal{U}_n\setminus k} \frac{x_l}{2}\ell_{n\leftarrow l}^{(i-1)} - \right.}{\sum_{\substack{\boldsymbol{x}^{[n]}\in\mathbb{X}^{d_n^f} \\ x_k=-1}} \exp\left(\sum_{l\in\mathcal{U}_n\setminus k} \frac{x_l}{2}\ell_{n\leftarrow l}^{(i-1)} - \right.} \frac{\left. -\frac{1}{2\sigma_{IN}^2}||y_n - \hat{\boldsymbol{h}}^{[n]T}\boldsymbol{x}^{[n]}||^2\right)}{\left. -\frac{1}{2\sigma_{IN}^2}||y_n - \hat{\boldsymbol{h}}^{[n]T}\boldsymbol{x}^{[n]}||^2\right)} \tag{29}$$

where $\boldsymbol{x}^{[n]}$ and $\hat{\boldsymbol{h}}^{[n]}$ are both d_n^f-length vectors, while $\hat{\boldsymbol{h}}^{[n]}$ contains the estimates to the channel of the users in the set of $\mathcal{U}_n$ related to the n-th AE.

Finally, when the affordable number of iterations is reached, the probability of the data symbol sent by the k-th user can be estimated as $P(x_k = \pm 1) = \eta_k \exp\left(\frac{x_k}{2}\sum_{n\in\mathcal{I}_k}\ell_{n\to k}\right)$, where η_k is selected to make $P(x_k=+1)+P(x_k=-1)=1$.

From (29) we can be inferred that the calculation of $\ell_{n\to k}^{(i)}$ can be executed separately by the individual AEs. The MPA-MUD can be implemented solely at the CPU or jointly between the CPU and AEs. Specifically for the joint implementation, the detection procedure can be carried out as follows. First, each AE estimates its related channels using the method provided in Section IV. At this channel estimation stage, the sets of $\mathcal{U}_n$ for $n = 1, \ldots, N$ can also be obtained with the aid of the OMP algorithm. Then, for $n = 1, \ldots, N$, the n-th AE transmits $\mathcal{U}_n$ to the CPU, based on which the factor graph can be constructed by the CPU. Then at the signal detection stage, the MPA-MUD can be jointly implemented at the CPU and AEs. In detail, during an iteration, each of the AEs first calculates the messages with the aid of (29), which are referred to as the FN messages.[4] Then, under the control of the CPU, AEs exchange their FN messages according to the connections defined by the factor graph. After obtaining the FN messages, each of the AEs updates the messages with the aid of (28), which are referred to as the VN messages. Then, the messages are exchanged according to the factor graph again with the support of the CPU. Then, the MPA-MUD forwards to the next iteration, and the above process repeats until the ending conditions are satisfied.

Note that, there is a certain probability that an AE only serves one user or several AEs only serve one user. In the first case, from (28) and (29) we can know that the detector carries out ML detection and there are no iterations needed. By contrast, in the second case, also from (28) and (29) we can know that the detector executes the distributed ML detection, where message exchange is needed, but only once.

From the above description we can conceive that in the MPA-MUD operated jointly by the CPU and AEs, most of the complexity is distributed among AEs, while the CPU only needs to maintain and update the factor graph, as well as control the exchange of messages. Therefore, the MPA-MUD operates the joint local-global detection. Most of the computations are distributed among AEs, which is in the principles of local signal detection. However, CPU is still needed to assist the signal detection, which is the contribution of global signal detection. By this way, using the complexity similar to local detection allows to attain the performance of near global detection. Furthermore, Backhaul resources are required for interchange of the intermediate messages among AEs. However, it is noteworthy that since the AEs serving a given user are usually neighbors, information exchange occurs between adjacent AEs with high probability. Therefore, in comparison with the global MUD, the required backhaul resources can be significantly reduced. In a little more detail, for global MUD, all the AEs in the system have to send all of their received signals to the CPU for carrying out the global signal detection. Hence, it puts forward an extreme demand on the throughput of the CPU, in particular, when a mDA system is large. By contrast, for the MPA-MUD, each of the AEs is only required to send the indices of its connected users to the CPU once per frame. Furthermore, as each AE is usually only connected with a very small number of users, the demand on the throughput of CPU can be significantly reduced in comparison with the global MUD. However, some extra backhaul resources are required by the MPA-MUD for interchanging the intermediate messages among AEs. Nevertheless, as mentioned in Section III-D considering the sparsity, a user is in general only served by a small number of AEs, for example, 3 AEs, when applying the parameters as shown in Section III-D. Consequently, the degrees of VN is small. Specifically, let d^v be the average degrees of VN. Then, we can be implied that for detecting one user's signal, only about d^v AEs are required to exchange their information.

VI. PERFORMANCE RESULTS

In our studies below, all distances are relative to a reference distance. The mDA systems are assumed to cover an area of 50×50. The two-slope propagation path-loss model of (4) has the parameters of $\nu_1 = 2$, $\nu_2 = 3$, and $g = 5$. For the lognormal shadowing effect, the standard deviation is set to 8 dB. Users and AEs are arranged according to the PPPs Φ_u and ϕ_a with the intensity of λ_u and λ_a, respectively. It is worth noting that in the following figures, the SNR per bit (E_b/N_0) represents *the transmit SNR, or more exactly, the transmit power per bit normalized by the noise power measured at AEs (receivers).* The reason for using transmit SNR instead of receive SNR is that our mDA system needs to simultaneously take into account of the large-scale (propagation path-loss and shadowing) and small-scale fading of individual users. Hence, even when the transmit SNR may be as high as 50 dB in our figures, the average SNR at receiver is actually very small after considering the propagation loss.[5]

[4] At the first iteration, $\ell_{n\leftarrow k}^{(0)}$ is initialized to zero for $n = 1, \cdots, N$ and $k = 1, \cdots, K$.

[5] In WiFi systems, the transmit power is usually at least 80 dB higher than the receive power.

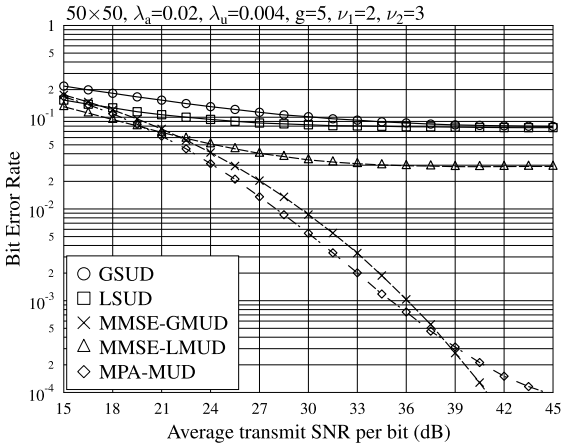


FIGURE 5. BER versus average transmit SNR per bit performance for the mDA systems employing various MUDs.

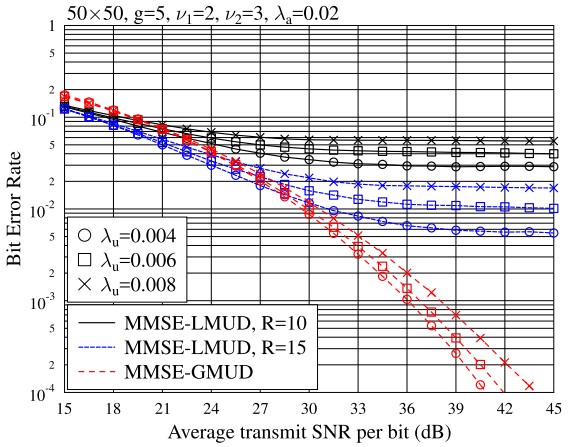


FIGURE 6. BER versus average transmit SNR per bit performance for the mDA systems having different densities of users, when the MMSE-based GMUD and LMUD are employed, respectively.

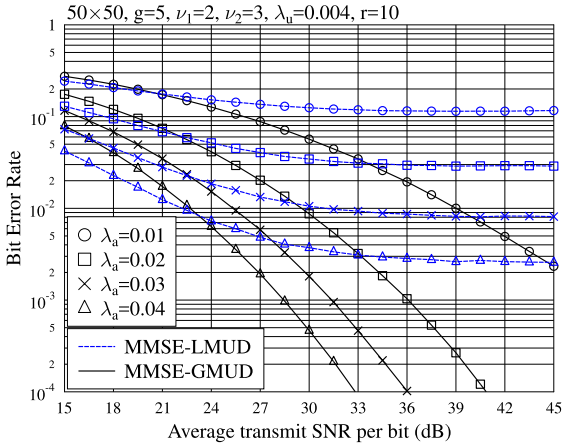


FIGURE 7. BER versus average transmit SNR per bit performance for the mDA systems deployed with different densities of AEs, when the MMSE-based GMUD and LMUD are employed, respectively.

Fig.5 compares the BER performance of the mDA systems employing various MUDs. Except the MPA-MUD that uses the CSI estimated from the OMP-based channel estimation, the global single user detection (GSUD), local SUD (LSUD), MMSE-GMUD and the MMSE-LMUD all adopt the CSI estimated from the LMMSE-based channel estimation. For the local detection schemes, we assume that a user is served by the AEs having their distance less than 10 from the user. In this case, a user can be served by approximately 6 AEs when $\lambda_a = 0.02$. Furthermore, when the MPA-MUD is considered, we assume that each AE serves no more than 4 users. As shown in Fig.5, although the average number of AEs is 50 and the average number of users is only 10, the BER performance of both GSUD and LSUD is very poor. This is because the effective number of AEs used by a user to suppress MUI is in general very small as the result of propagation path-loss, which is insufficient for the MUI suppression in the correlation principle. Furthermore, it can be observed from Fig.5 that the BER performance of the GSUD is worse than that of the LSUD, which is the result of the channel estimation error. When the GSUD is considered for detecting one user, all AEs have to estimate the channels from the user, many of them are estimated based on very weak signals due to propagation pass-loss, which are hence very inaccurate. The even worse case is that there are probably other relatively strong users close to these AEs, making the channel estimation for weak users extremely difficult. Consequently, the signals provided by the AEs for a weak user contribute only interference, instead of its performance enhancement. This phenomenon can also be observed between the MMSE-GMUD and MMSE-LMUD in low SNR region, where channel estimation is less reliable. However, when SNR increases and hence the channel estimation of all AEs becomes sufficiently reliable, MMSE-GMUD is able to achieve better BER performance than MMSE-LMUD, and even than MPA-MUD, if SNR is sufficiently high. Finally, the proposed MPA-MUD is capable of attaining the best BER performance than all the other detectors considered in the main SNR region. However, for Fig.5, the constraint on the detection complexity is imposed on MPA-MUD, which assumes that each AE serves no more than 4 users, while the signals from the other users are treated as Gaussian noise. Consequently, error floor may be observed in high SNR region, as shown in Fig.5.

Fig.6 and Fig.7 illustrate the impact of user density and AE density on the achievable BER performance of the mDA systems employing respectively the MMSE-based GMUD and LMUD. Specifically, in Fig.6, two cases for the MMSE-LMUD are considered, which assume that a user is served by the AEs having their distances less than $R = 10$ and $R = 15$ from the user. Furthermore, we note that the MMSE-GMUD is equivalent to a MMSE-LMUD with $R \to \infty$. Explicitly, Fig.6 shows that when R becomes larger, i.e., when a user is in average served by more AEs, the BER performance improves. Correspondingly, the complexity demanded also increases, as R increases. Fig.6 also shows that for a given case, the BER performance of mDA systems degrades, as the user density increases, due to the increased MUI. By contrast, in Fig.7, we consider the effect of AE density, when $R = 10$ is assumed. Explicitly, the improvement of BER performance is significant with the

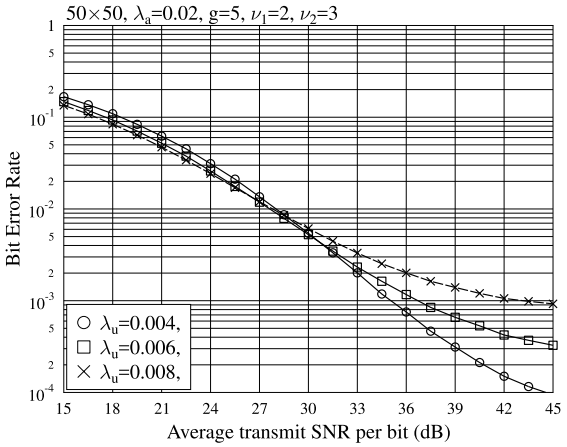


FIGURE 8. BER versus average transmit SNR per bit performance for the mDA systems supporting different densities of users, when the OMP-based channel estimation and MPA-MUD are employed.

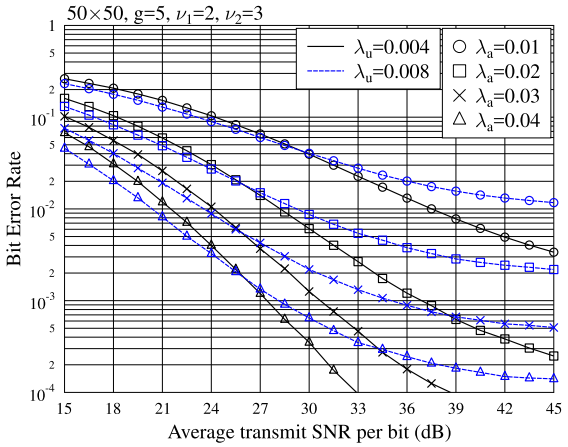


FIGURE 9. BER versus average transmit SNR per bit performance for the mDA systems with different AE and user densities, when the OMP-based channel estimation and MPA-MUD are employed.

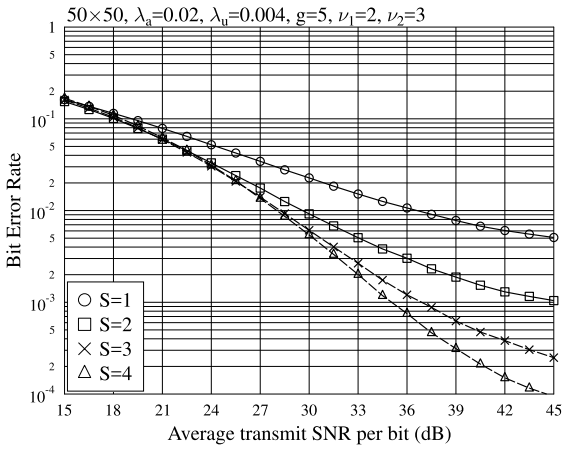


FIGURE 10. BER versus average transmit SNR per bit performance for the mDA systems, where different number of users is managed by each of the AEs, and when the OMP-based channel estimation and MPA-MUD are employed.

increase of λ_a. This is because increasing the AE density yields twofold of positive effect. First, for a given user and a given value of R, increasing the AE density provides a user the opportunity of being served by more AEs, and hence increases the diversity gain and consequently, improves the BER performance mainly in the high-SNR regime. Second, when the AE density increases, the distances between users and AEs in general reduce, which yields the power gain resulting in performance improvement noticeably in the low-SNR regime. Owing to the above-mentioned twofold positive effect, the overall BER performance of mDA systems improves with the increase of λ_a.

Fig.8 shows the BER performance of the mDA systems having different intensity of users, i.e., supporting different number of users, when the OMP-based channel estimation and MPA-MUD are employed. Furthermore, it is assumed that any of the AEs is able to serve 4 users, who have the largest estimated channel gains by the AE. The other parameters used in simulation are detailed on the top of the figure. From Fig.8 it can be observed that in low SNR region, more users supported by the mDA system results in slight improvement of the BER performance. This is the result of the multiuser diversity gain provided by the OMP-based channel estimation. By contrast, in high SNR region, the BER performance of mDA systems degrades, as the number of users increases, due to the fact that MUI dominates the achievable BER performance in this case.

Fig.9 shows the BER performance of the mDA systems with different AE and user intensities, when the OMP-based channel estimation and MPA-MUD are employed. For the results, it is assumed that an AE is able to serve 3 users with the largest channel gains identified by the AE. From Fig.9 we observe that with the increase of λ_a, the BER performance of mDA systems improves significantly, as explained in association with Fig.7. However, error floors are observed in Fig.9 in all the cases, except the one of $\lambda_a = 0.04$ and $\lambda_u = 0.004$. In this case, the number of AEs is 10 times of the number of users, and hence MUI can be near-ideally suppressed by the receiver.

Finally, in Fig.10 and Table 1, we demonstrate the impact of the number of users, expressed as S, allowed to be managed by an AE on the BER performance of the mDA systems, when the parameters used are shown with the figure. Again, the served users by an AE are the S strongest users estimated by the AE using the proposed OMP-based channel estimation. Note that, as described in Algorithm 1, S is actually the number of iterations. It can be seen from Fig.10 that, when the number of users manageable by an AE is increased from $S = 1$ to $S = 4$, the BER performance of mDA systems improves. Certainly, the cost for this performance improvement is the increase of AE's complexity. However, Table 1 shows that if an AE is forced to handle many users, the BER performance will degrade, which is resulted from the channel estimation errors, as in this case an AE has to estimate some weak users. Therefore, as shown in Table 1, for a given SNR, there is an optimal value for S, which yields the lowest BER. Additionally, as SNR increases, the optimal value of S also increases. This is because higher SNR leads to low channel estimation error. Consequently, an AE is able to serve a larger range with required reliability, allowing a larger value for the optimum S.

TABLE 1. BER performance comparison when different number of users is managed at each of the AEs.

SNR per bit	$S=1$	$S=2$	$S=3$	$S=4$	$S=5$	$S=6$	$S=7$
15dB	0.161135	**0.154662**	0.160932	0.167989	0.172592	0.176275	0.179492
30dB	0.022660	0.009278	0.006141	**0.005397**	0.005585	0.006102	0.007392
45dB	5.0250E-03	1.0153E-03	2.4570E-04	**9.4812E-05**	1.2251E-04	3.2497E-04	1.0092E-03

VII. CONCLUSION

The uplink signal detection in mDA systems has been investigated, where the distributions of both AEs and users are assumed to follow the PPP model with the density of AEs being higher than the users' density. We have firstly provided the insightful analysis to show that in mDA systems, local detection is more appropriate and has more advantages than global detection. We have conceived that the channels of a mDA system are sparse nature. Hence, the OMP-based channel estimation has been introduced for both acquiring CSI and identifying the users associated with different AEs. Furthermore, based on the association of users with AEs, a factor graph can be built for operation of the proposed MPA-MUD. It can be shown that the MPA-MUD is a low-complexity MUD, which can be solely implemented at the CPU, or jointly operated between the CPU and AEs. In the later case, most of the computations can be distributed among the AEs, while the CPU is only required to maintain the factor graph to control the information exchange between AEs. Our simulation results show that the proposed MPA-MUD can significantly outperform the local detection, which has low complexity. Meanwhile, within the SNR region of practical interest, the performance achieved by the proposed MPA-MUD is close to that of the global detection, which is high-complexity and requires huge resources for information exchange. Overall, our studies show that the OMP-based channel estimation and the MPA-MUD can be employed to gain a good trade-off among BER performance, complexity and resource usage.

APPENDIX A
SOME LAPLACE TRANSFORMS USED IN THE PAPER

Let $I=\sum_{n=1}^{N}\beta_{n,1}\alpha_{1,n}^2$ denote the power received by N AEs from one transmitting user, when the transmitted power of the user is normalized to 1. The LT of I is given by $\mathbb{E}\left[e^{-sI}\right]$. Below we consider three cases. The first case assumes that I is the power received by all the AEs in a mDA system covering an infinite area. In this case, the total received power I is denoted as I_{all}. In the second case, the power received by the AEs having their distances less than R from the user is considered. Correspondingly, the received power I is denoted by I_R. Finally, the third case addresses the power received by the AEs having their distances larger than R from the user, and the received power I in this case is expressed as $\bar{I}_R$. Obviously, we have $I_{\text{all}}=I_R+\bar{I}_R$.

In $I=\sum_{n=1}^{N}\beta_{n,1}\alpha_{1,n}^2$, $\beta_{n,1}$ is depended on the reference distance r, and $\alpha_{1,n}^2$ obeys the exponential distribution. Hence, with the aid of the probability generating functional [44], and by mapping the PPP to one dimension representation, the LT of I can be expressed as [44]

$$\mathcal{L}(s)=\exp\left[-2\lambda_a\pi\int_{r_1}^{r_2}\mathbb{E}_{\alpha^2}\left(1-e^{-s\alpha^2\beta(x)}\right)xdx\right] \quad (30)$$

where α^2 is a variable obeying the exponential distribution, and $\beta(x)=x^{-\nu}$ when $x>1$, and otherwise, $\beta(x)=1$ when $x\leq 1$. Furthermore, in (30), we have $r_1=0$ and $r_2=\infty$ for the above noted first case, $r_1=0$ and $r_2=R$ for the second case, and $r_1=R$ and $r_2=\infty$ for the third case.

Applying the exponential distribution of α^2, (30) can be rewritten as

$$\begin{aligned}\mathcal{L}(s)&=\exp\left[-2\lambda_a\pi\int_0^{\infty}\int_{r_1}^{r_2}\left(1-e^{-sy\beta(x)}\right)xe^{-y}dxdy\right]\\&\triangleq\exp(\Lambda(s))\end{aligned} \quad (31)$$

Therefore, with the aid of some mathematical simplification tools, such as, [48], in the first case, it can be shown that

$$\begin{aligned}\Lambda(s)&=-2\lambda_a\pi\left[\int_0^{\infty}\int_0^{1}\left(1-e^{-sy}\right)xe^{-y}dxdy\right.\\&\quad\left.+\int_0^{\infty}\int_1^{\infty}\left(1-e^{-syx^{-\nu}}\right)xe^{-y}dxdy\right]\\&=\frac{\lambda_a\pi}{s+1}-s^{\delta}\frac{\lambda_a\pi}{\text{sinc}\delta}-\frac{\lambda_a\pi\delta}{(\delta+1)(s+1)}\\&\quad\times{}_2F_1\left(1,1;\delta+2;\frac{1}{s+1}\right)\end{aligned} \quad (32)$$

where $\delta=2/\nu$. In the second case, we have

$$\begin{aligned}\Lambda(s)&=-2\lambda_a\pi\left[\int_0^{\infty}\int_0^{1}\left(1-e^{-sy}\right)xe^{-y}dxdy\right.\\&\quad\left.+\int_0^{\infty}\int_1^{R}\left(1-e^{-syx^{-\nu}}\right)xe^{-y}dxdy\right]\\&=\frac{\lambda_a\pi}{s+1}-\frac{\lambda_a\pi\delta}{(\delta+1)(s+1)}\,{}_2F_1\left(1,1;\delta+2;\frac{1}{s+1}\right)\\&\quad+\frac{\lambda_a\pi\delta R^2}{(\delta+1)(sR^{-\nu}+1)}\,{}_2F_1\left(1,1;\delta+2;\frac{1}{sR^{-\nu}+1}\right)\\&\quad-\lambda_a\pi R^2\end{aligned} \quad (33)$$

when $R>1$. Finally, in the third case, since $I_{\text{all}}=I_R+\bar{I}_R$, the LT of $\bar{I}_R$, denoted by $\mathcal{L}_{\bar{I}_R}(s)$, can be obtained as $\mathcal{L}_{\bar{I}_R}(s)=\mathcal{L}_{I_{\text{all}}}(s)/\mathcal{L}_{I_R}(s)$, when $R>1$.

REFERENCES

[1] M. Shafi, A. F. Molisch, P. J. Smith, T. Haustein, P. Zhu, P. De Silva, F. Tufvesson, A. Benjebbour, and G. Wunder, "5G: A tutorial overview of standards, trials, challenges, deployment, and practice," *IEEE J. Sel. Areas Commun.*, vol. 35, no. 6, pp. 1201–1221, Jun. 2017.
[2] S. Chen, F. Qin, B. Hu, X. Li, and Z. Chen, "User-centric ultra-dense networks for 5G: Challenges, methodologies, and directions," *IEEE Wireless Commun.*, vol. 23, no. 2, pp. 78–85, Apr. 2016.

[3] M. Kamel, W. Hamouda, and A. Youssef, "Ultra-dense networks: A survey," *IEEE Commun. Surveys Tuts.*, vol. 18, no. 4, pp. 2522–2545, 4th Quart., 2016.

[4] H. Zhu, "Performance comparison between distributed antenna and microcellular systems," *IEEE J. Sel. Areas Commun.*, vol. 29, no. 6, pp. 1151–1163, Jun. 2011.

[5] D. Gesbert, S. Hanly, H. Huang, S. Shamai (Shitz), O. Simeone, and W. Yu, "Multi-cell MIMO cooperative networks: A new look at interference," *IEEE J. Sel. Areas Commun.*, vol. 28, no. 9, pp. 1380–1408, Dec. 2010.

[6] R. Irmer, H. Droste, P. Marsch, M. Grieger, G. Fettweis, S. Brueck, H.-P. Mayer, L. Thiele, and V. Jungnickel, "Coordinated multipoint: Concepts, performance, and field trial results," *IEEE Commun. Mag.*, vol. 49, no. 2, pp. 102–111, Feb. 2011.

[7] P. Pan, Y. Zhang, L.-L. Yang, and X. Ju, "Capacity of generalised network multiple-input–multiple-output systems with multicell cooperation," *IET Commun.*, vol. 7, no. 17, pp. 1925–1937, Nov. 2013.

[8] G. Zhai, L. Tian, Y. Zhou, Q. Sun, and J. Shi, "A computing resource adjustment mechanism for communication protocol processing in centralized radio access networks," *China Commun.*, vol. 13, no. 12, pp. 79–89, Dec. 2016.

[9] G. Zhai, L. Tian, Y. Zhou, and J. Shi, "Load diversity based processing resource allocation for super base stations in large-scale centralized radio access networks," in *Proc. IEEE Int. Conf. Commun. (ICC)*, Jun. 2014, pp. 5119–5124.

[10] C. Pan, M. Elkashlan, J. Wang, J. Yuan, and L. Hanzo, "User-centric C-RAN architecture for ultra-dense 5G networks: Challenges and methodologies," *IEEE Commun. Mag.*, vol. 56, no. 6, pp. 14–20, Jun. 2018.

[11] A. Checko, H. L. Christiansen, Y. Yan, L. Scolari, G. Kardaras, M. S. Berger, and L. Dittmann, "Cloud RAN for mobile networks—A technology overview," *IEEE Commun. Surveys Tuts.*, vol. 17, no. 1, pp. 405–426, 1st Quart., 2015.

[12] 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.

[13] E. Bjornson 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.

[14] J. Qiu, K. Xu, X. Xia, Z. Shen, and W. Xie, "Downlink power optimization for cell-free massive MIMO over spatially correlated Rayleigh fading channels," *IEEE Access*, vol. 8, pp. 56214–56227, Mar. 2020.

[15] A. Abdallah and M. M. Mansour, "Efficient angle-domain processing for FDD-based cell-free massive MIMO systems," *IEEE Trans. Commun.*, vol. 68, no. 4, pp. 2188–2203, Apr. 2020.

[16] J. Li, D. Wang, P. Zhu, J. Wang, and X. You, "Downlink spectral efficiency of distributed massive MIMO systems with linear beamforming under pilot contamination," *IEEE Trans. Veh. Technol.*, vol. 67, no. 2, pp. 1130–1145, Feb. 2018.

[17] F. Yuan, S. Jin, Y. Huang, K.-K. Wong, Q. T. Zhang, and H. Zhu, "Joint wireless information and energy transfer in massive distributed antenna systems," *IEEE Commun. Mag.*, vol. 53, no. 6, pp. 109–116, Jun. 2015.

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

[19] G. N. Kamga, M. Xia, and S. Aissa, "Spectral-efficiency analysis of massive MIMO systems in centralized and distributed schemes," *IEEE Trans. Commun.*, vol. 64, no. 5, pp. 1930–1941, May 2016.

[20] J. Li, D. Wang, P. Zhu, and X. You, "Uplink spectral efficiency analysis of distributed massive MIMO with channel impairments," *IEEE Access*, vol. 5, pp. 5020–5030, 2017.

[21] J. Li, D. Wang, X. You, and P. Zhu, "Spectral efficiency analysis of single-cell multi-user large-scale distributed antenna system," *IET Commun.*, vol. 8, no. 12, pp. 2213–2221, Aug. 2014.

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

[23] S. Govindasamy and I. Bergel, "Uplink performance of multi-antenna cellular networks with co-operative base stations and user-centric clustering," *IEEE Trans. Wireless Commun.*, vol. 17, no. 4, pp. 2703–2717, Apr. 2018.

[24] J. Yuan, S. Jin, W. Xu, W. Tan, M. Matthaiou, and K.-K. Wong, "User-centric networking for dense C-RANs: High-SNR capacity analysis and antenna selection," *IEEE Trans. Commun.*, vol. 65, no. 11, pp. 5067–5080, Nov. 2017.

[25] H. Q. Ngo, L.-N. Tran, T. Q. Duong, M. Matthaiou, and E. G. Larsson, "On the total energy efficiency of cell-free massive MIMO," *IEEE Trans. Green Commun. Netw.*, vol. 2, no. 1, pp. 25–39, Mar. 2018.

[26] J. Zuo, J. Zhang, C. Yuen, W. Jiang, and W. Luo, "Energy-efficient downlink transmission for multicell massive DAS with pilot contamination," *IEEE Trans. Veh. Technol.*, vol. 66, no. 2, pp. 1209–1221, Feb. 2017.

[27] R. Senanayake, P. Smith, P. L. Yeoh, and J. Evans, "An SNR approximation for distributed massive MIMO with zero forcing," *IEEE Commun. Lett.*, vol. 19, no. 11, pp. 1885–1888, Nov. 2015.

[28] H. Yin, D. Gesbert, and L. Cottatellucci, "Dealing with interference in distributed large-scale MIMO systems: A statistical approach," *IEEE J. Sel. Topics Signal Process.*, vol. 8, no. 5, pp. 942–953, Oct. 2014.

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

[30] C. Pan, H. Zhu, N. J. Gomes, and J. Wang, "Joint user selection and energy minimization for ultra-dense multi-channel C-RAN with incomplete CSI," *IEEE J. Sel. Areas Commun.*, vol. 35, no. 8, pp. 1809–1824, Aug. 2017.

[31] T. X. Tran and D. Pompili, "Dynamic radio cooperation for user-centric cloud-RAN with computing resource sharing," *IEEE Trans. Wireless Commun.*, vol. 16, no. 4, pp. 2379–2393, Apr. 2017.

[32] Z. Zhang, N. Wang, J. Zhang, and X. Mu, "Dynamic user-centric clustering for uplink cooperation in multi-cell wireless networks," *IEEE Access*, vol. 6, pp. 8526–8538, 2018.

[33] Y. Liu, X. Li, F. R. Yu, H. Ji, H. Zhang, and V. C. M. Leung, "Grouping and cooperating among access points in user-centric ultra-dense networks with non-orthogonal multiple access," *IEEE J. Sel. Areas Commun.*, vol. 35, no. 10, pp. 2295–2311, Oct. 2017.

[34] C. Pan, H. Zhu, N. J. Gomes, and J. Wang, "Joint precoding and RRH selection for user-centric green MIMO C-RAN," *IEEE Trans. Wireless Commun.*, vol. 16, no. 5, pp. 2891–2906, May 2017.

[35] A. Liu and V. K. N. Lau, "Two-timescale user-centric RRH clustering and precoding optimization for cloud RAN via local stochastic cutting plane," *IEEE Trans. Signal Process.*, vol. 66, no. 1, pp. 64–76, Jan. 2018.

[36] Y. Zhang, S. Bi, and Y.-J.-A. Zhang, "User-centric joint transmission in virtual-cell-based ultra-dense networks," *IEEE Trans. Veh. Technol.*, vol. 67, no. 5, pp. 4640–4644, May 2018.

[37] E. Nayebi, A. Ashikhmin, T. L. Marzetta, H. Yang, and B. D. Rao, "Precoding and power optimization in cell-free massive MIMO systems," *IEEE Trans. Wireless Commun.*, vol. 16, no. 7, pp. 4445–4459, Jul. 2017.

[38] C. Pan, H. Mehrpouyan, Y. Liu, M. Elkashlan, and N. Arumugam, "Joint pilot allocation and robust transmission design for ultra-dense user-centric TDD C-RAN with imperfect CSI," *IEEE Trans. Wireless Commun.*, vol. 17, no. 3, pp. 2038–2053, Mar. 2018.

[39] C.-K. Wen, J.-C. Chen, K.-K. Wong, and P. Ting, "Message passing algorithm for distributed downlink regularized zero-forcing beamforming with cooperative base stations," *IEEE Trans. Wireless Commun.*, vol. 13, no. 5, pp. 2920–2930, May 2014.

[40] Y. Lin, R. Zhang, L. Yang, and L. Hanzo, "Secure user-centric clustering for energy efficient ultra-dense networks: Design and optimization," *IEEE J. Sel. Areas Commun.*, vol. 36, no. 7, pp. 1609–1621, Jul. 2018.

[41] T. M. Hoang, H. Q. Ngo, T. Q. Duong, H. D. Tuan, and A. Marshall, "Cell-free massive MIMO networks: Optimal power control against active eavesdropping," *IEEE Trans. Commun.*, vol. 66, no. 10, pp. 4724–4737, Oct. 2018.

[42] T. A. Khan, P. V. Orlik, K. J. Kim, R. W. Heath, and K. Sawa, "A stochastic geometry analysis of large-scale cooperative wireless networks powered by energy harvesting," *IEEE Trans. Commun.*, vol. 65, no. 8, pp. 3343–3358, Aug. 2017.

[43] L.-L. Yang and W. Fang, "Performance of distributed-antenna DS-CDMA systems over composite lognormal shadowing and Nakagami-*m*-fading channels," *IEEE Trans. Veh. Technol.*, vol. 58, no. 6, pp. 2872–2883, Jul. 2009.

[44] M. Haenggi, *Stochastic Geometry for Wireless Networks*. New York, NY, USA: Cambridge Univ. Press, 2013.

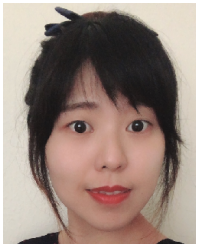
[45] M. K. Simon and M.-S. Alouini, *Digital Communication Over Fading Channels: A Unified Approach to Performance Analysis*. New York, NY, USA: Wiley, 2000.

[46] P. Harley, "Short distance attenuation measurements at 900 MHz and 1.8 GHz using low antenna heights for microcells," *IEEE J. Sel. Areas Commun.*, vol. 7, no. 1, pp. 5–11, Jan. 1989.

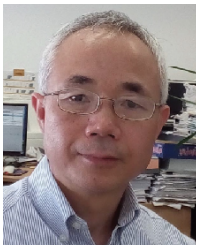
[47] J. G. Andrews, F. Baccelli, and R. K. Ganti, "A tractable approach to coverage and rate in cellular networks," *IEEE Trans. Commun.*, vol. 59, no. 11, pp. 3122–3134, Nov. 2011.
[48] I. S. Gradshteyn and I. M. Ryzhik, *Table of Integrals, Series, and Products*, 7th ed. New York, NY, USA: Academic, 2007.
[49] M. K. Simon and M.-S. Alouini, *Digital Communication Over Fading Channels: A Unified Approach to Performance Analysis*. New York, NY, USA: Wiley, 2000.
[50] S. J. Wright, R. D. Nowak, and M. A. T. Figueiredo, "Sparse reconstruction by separable approximation," *IEEE Trans. Signal Process.*, vol. 57, no. 7, pp. 2479–2493, Jul. 2009.
[51] M. Biguesh and A. B. Gershman, "Training-based MIMO channel estimation: A study of estimator tradeoffs and optimal training signals," *IEEE Trans. Signal Process.*, vol. 54, no. 3, pp. 884–893, Mar. 2006.
[52] J. Wang and P. Li, "Recovery of sparse signals using multiple orthogonal least squares," *IEEE Trans. Signal Process.*, vol. 65, no. 8, pp. 2049–2062, Apr. 2017.
[53] J. A. Tropp and A. C. Gilbert, "Signal recovery from random measurements via orthogonal matching pursuit," *IEEE Trans. Inf. Theory*, vol. 53, no. 12, pp. 4655–4666, Dec. 2007.
[54] T. J. Richardson and R. L. Urbanke, "The capacity of low-density parity-check codes under message-passing decoding," *IEEE Trans. Inf. Theory*, vol. 47, no. 2, pp. 599–618, Feb. 2001.
[55] G. Dziwoki, "Averaged properties of the residual error in sparse signal reconstruction," *IEEE Signal Process. Lett.*, vol. 23, no. 9, pp. 1170–1173, Sep. 2016.
[56] H. Wang, W. Zhang, Y. Liu, Q. Xu, and P. Pan, "On design of non-orthogonal pilot signals for a multi-cell massive MIMO system," *IEEE Wireless Commun. Lett.*, vol. 4, no. 2, pp. 129–132, Apr. 2015.
[57] J. I. Haas and P. G. Casazza, "On the structures of Grassmannian frames," in *Proc. Int. Conf. Sampling Theory Appl. (SampTA)*, Jul. 2017, pp. 377–380.
[58] Z. Chen, X. Hou, and C. Yang, "Training resource allocation for user-centric base station cooperation networks," *IEEE Trans. Veh. Technol.*, vol. 65, no. 4, pp. 2729–2735, Apr. 2016.
[59] H. Loeliger, "An introduction to factor graphs," *IEEE Signal Process. Mag.*, vol. 21, no. 1, pp. 28–41, Jan. 2004.
[60] A. Bayesteh, H. Nikopour, M. Taherzadeh, H. Baligh, and J. Ma, "Low complexity techniques for SCMA detection," in *Proc. IEEE Globecom Workshops (GC Wkshps)*, Dec. 2015, pp. 1–6.

PENG PAN (Senior Member, IEEE) received the B.Eng. and Ph.D. degrees in electronic engineering from Beihang University, Beijing, China, in 2005 and 2011, respectively.

From December 2007 to December 2008, he was a Visiting Ph.D. Student with the School of Electronics and Computer Science, The University of Southampton, Southampton, U.K. He is currently an Associate Professor with the School of Communication Engineering, Hangzhou Dianzi University, Hangzhou, China. His research interests include MIMO systems, multiuser detection, and performance evaluation of wireless systems.



JIATIAN ZHANG (Graduate Student Member, IEEE) received the M.Sc. degree in electrical and telecommunications engineering from Shandong University, Jinan, China, in 2012. She is currently pursuing the Ph.D. degree with the Next Generation of Wireless Group, The University of Southampton. From 2013 to 2015, she worked with the Operation Department, ZTE Corporation as a Senior Engineer across many countries globally. From 2015 to 2017, she worked as the Pre-5G Project Wireless Solution Manager with ZTE Japan. Her current research interests include grant-free multiple access for massive MIMO networks and distributed massive MIMO systems.



LIE-LIANG YANG (Fellow, IEEE) received the B.Eng. degree in communications engineering from Shanghai Tiedao University, Shanghai, China, in 1988, and the M.Eng. and Ph.D. degrees in communications and electronics from Northern (Beijing) Jiaotong University, Beijing, China, in 1991 and 1997, respectively. From June 1997 to December 1997, he was a Visiting Scientist with the Institute of Radio Engineering and Electronics, Academy of Sciences of the Czech Republic. Since December 1997, he has been with The University of Southampton, U.K., where he is currently a Professor of wireless communications with the School of Electronics and Computer Science. His research interests include wireless communications, wireless networks and signal processing for wireless communications and molecular communications, and nano-networks. He has published over 390 research articles in journals and conference proceedings, authored/coauthored three books, and also published several book chapters. He is a Fellow of IET, and was a Distinguished Lecturer of the IEEE VTS. He served as an Associate Editor to the IEEE TRANSACTIONS ON VEHICULAR TECHNOLOGY and *Journal of Communications and Networks* (JCN). He is an Associate Editor to IEEE ACCESS and a Subject Editor to the *Electronics Letters*.