Coded Pilot Random Access for Massive MIMO Systems

Sørensen, Jesper Hemming; Carvalho, Elisabeth De; Stefanovic, Cedomir; Popovski, Petar

Published in:
I E E E Transactions on Wireless Communications

DOI (link to publication from Publisher):
10.1109/TWC.2018.2873400

Publication date:
2018

Document Version
Accepted author manuscript, peer reviewed version

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TWC.2018.2873400, IEEE Transactions on Wireless Communications

Coded Pilot Random Access for Massive MIMO Systems

Jesper H. Sørensen, Elisabeth de Carvalho, Čedomir Stefanović and Petar Popovski
Aalborg University, Department of Electronic Systems, Fredrik Bajers Vej 7, 9220 Aalborg, Denmark
E-mail: {jhs,edc,cs,petarp}@es.aau.dk

***Abstract*—We present a novel access protocol for crowd scenarios in massive MIMO (Multiple-input multiple-output) systems. Crowd scenarios are characterized by a large number of users with intermittent access behavior, whereby orthogonal scheduling is infeasible. In such scenarios, random access is a natural choice. The proposed access protocol relies on two essential properties of a massive MIMO system, namely asymptotic orthogonality between user channels and asymptotic invariance of channel powers. Signal processing techniques that take advantage of these properties allow us to view a set of contaminated pilot signals as a graph code on which iterative belief propagation can be performed. This makes it possible to decontaminate pilot signals and increase the throughput of the system. Numerical evaluations show that the proposed access protocol increases the throughput with 36%, when having 400 antennas at the base station, compared to the conventional method of slotted ALOHA. With 1024 antennas, the throughput is increased by 85%.**

I. Introduction

Massive MIMO (Multiple-input multiple-output) has been identified as a key technology to improve spectral efficiency of wireless communication systems and one of the main enablers of the upcoming 5th generation [1]. A massive MIMO system refers to a multi-cell multi-user system with a massive number of antennas at the BS that serves multiple users [2]. The number of users is much smaller than the number of BS antennas, defining an under-determined multi-user system with a massive number of extra spatial degrees of freedom (DoF). Exploiting those extra DoF and assuming an infinite number of antennas at the BS, the multi-user MIMO channel can be turned into an orthogonal channel and the effects of small-scale fading and thermal noise can be eliminated.

However, when the number of antennas becomes massive, acquiring the channel state information (CSI) becomes a severe bottleneck. Downlink channel training requires a training length that is proportional to the number of antennas at the BS and is thus impractical. A solution promoted in [2] restricts massive MIMO operations to time-division duplex (TDD) for which channel reciprocity is exploited. As the downlink and uplink channels are equal, CSI is acquired at the BS based on uplink training and then used for downlink transmission. The benefit is that the training length is proportional to the number of users, which is much smaller than the number of BS antennas. Users in the same cell are assigned orthogonal pilot sequences, but, due to the shortage of orthogonal sequences, the same pilot sequences must be reused in neighboring cells, causing pilot contamination. Hence, pilot contamination is usually seen as an inter-cell interference problem. Various approaches exist to mitigate pilot contamination which can be classified according to increased inter-cell coordination requirements. Within the first class of approaches, not requiring inter-cell coordination, [3]–[6] utilizes semi-blind estimation to separate the subspace occupied by the channel of the desired user from the subspace occupied by the channels of the interfering users. In [7], allocation of pilot sequences is randomized across uplink transmission time slots so that the effect of pilot decontamination gets averaged out over time. One second approach relies on a coordinated pilot reuse plan across cells where users occupying different angular domains seen from a given base station can reuse the same pilot sequences [8]–[10]. The third approach assumes a high level inter-cell coordination where the data and CSI are shared at all cooperating BSs. A joint coordinated processing is performed that rejects the interfering signals created by pilot contamination [11], [12]. Lastly, new results [13] show that the spectral efficiency of massive MIMO does not saturate due to pilot contamination when the number of antennas grows asymptotically large, if the spatial correlation matrices of the pilot contaminating users are asymptotically linearly independent (not necessarily orthogonal) and optimal linear processing is employed.

In all the methods previously cited, an implicit assumption is that the pilot sequences of the users associated with the same cell are perfectly scheduled, such that no intra-cell pilot contamination occurs. These assumptions fall apart when one considers very dense, crowd scenarios as envisioned in 5G wireless scenarios [14]. In such a setting, orthogonal scheduling of the users belonging to the same BS becomes infeasible due to scheduling overhead. As an example, crowd sensor networks are characterized by intermittent and random activity, which makes the scheduling overhead prohibitive. Furthermore, human-oriented non-streaming internet traffic is intermittent and thus subject to significant scheduling overhead. This work is motivated by the massive MIMO problem in a crowd setting, as well as the observation that the pilot contamination problem is very much dependent on the *protocol assumptions* made in the system. Specifically, we consider a crowd scenario where the amount of users

The research presented in this paper was partially supported by the Danish Council for Independent Research (Det Frie Forskningsråd), grants no. DFF-1335-00273 and DFF-4005-00281. Part of this work has been performed in the framework of the Horizon 2020 project FANTASTIC-5G (ICT-671660), which is partly funded by the European Union. The authors would like to acknowledge the contributions of their colleagues in FANTASTIC-5G.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TWC.2018.2873400, IEEE Transactions on Wireless Communications

and their access behavior make it infeasible to schedule the transmissions. Instead users choose pilot sequences at random in an uncoordinated manner from a small pool shared by all users. In this way, the inter-cell pilot contamination problem becomes an intra-cell pilot contamination problem, where the BS needs to handle collisions that occur in the pilot domain.

The general approach devised in this paper employs a random access procedure to the pilot sequences. Random access is used when the number of potential resource users is larger than the actual number of resources, but not all potential users are simultaneously active. In our work we have created such a setup for access to the pilot sequences. It is important, however, to make the distinction between the random access procedure for link establishment used for example in LTE [15] and the random access procedure described in this paper. In the first procedure, a user wishing to establish a connection to the base station selects a preamble at random. If it enters a collision with a user which has selected the same preamble, it reiterates the process, otherwise it starts exchanging control information with the base station to establish the connection, meaning identify itself and provide useful information about the link quality. A vast literature exists on the general topic of random access for link establishment and its variants. More directly related to our work, random access protocols for massive machine-type communications are presented in [16]. For a massive MIMO system, a method for collision resolution is described in [17] exploiting the asynchronicity in the preambles used for network access. The random access procedure to pilots is performed *after* the connection is established and the random access is performed on the set of pilot sequences in order to estimate the channels of the users. To avoid any confusion, we use the denomination *Random Access to Pilots (RAP)* introduced in [18] to generally refer to random access procedures to pilots.

RAP can be implemented following different kinds of random access protocols [18]. RAP was initially considered in [19] in a massive MIMO system. [19] is the basis of the present paper and proposes a joint pilot and data transmission in the uplink according to a *coded random access protocol* [20], [21]. Likewise, the paper [22] describes a time-slotted pilot and data transmission, where, in each time slot, a user selects uniformly at random a pilot sequence and selects part of its codeword. The packet collisions are neither detected nor resolved, while over an asymptotically long time horizon, fading and effects of pilot contamination are averaged out allowing the determination of a reliable rate for transmission. Another related work is [23], where the pilots are transmitted by using the random access procedure in LTE, but modified according to the specifics of pilot access. The paper proposes an approach to resolve one-shot collisions by exploiting the channel hardening properties of massive MIMO and enabling the terminals to detect the collision and act accordingly. An extension of [23] can be found in [24], [25]. The paper [26] presents a different approach where the users are assigned unique non-orthogonal pilot sequences. User identification based on the pilot sequences and channel estimation is carried out employing compressed sensing techniques.

As previously mentioned, the protocol in this paper follows a *coded random access* procedure [20], [21]. The terminals transmit with a predefined probability and send a pilot in the uplink followed by the data part. Pilot assignment is randomized in each time slot while the data part is repeated, which enables to use successive interference cancellation (SIC) across the replicas of the same packet. The proposed implementation of SIC is a novel utilization of the massive MIMO properties, as it relies on two features specific to this context: (1) asymptotic orthogonality between user channels; and (2) asymptotic invariance of the power received from a user over a short time interval. These properties allow a processing of the received signals, which turns contaminated signals into linear combinations of data. These linear combinations form a codeword, which can be decoded using SIC. The ability to decontaminate signals through SIC provides an increase in throughput compared to conventional methods, like framed slotted ALOHA, which is considered the reference in this work. In framed slotted ALOHA, the terminals operate like in the proposed scheme, however the probability of transmission is optimized for contamination free transmissions, since SIC is not available.

In summary, the main contributions of this work are: (1) A novel coded random access framework specifically designed for a massive MIMO system; and (2) analysis using and-or tree evaluation with modifications accounting for the physical layer aspects of the proposed scheme. Compared to the preliminary version of this work in [19], the present version provides a thorough elaboration on the and-or tree analysis of the degree distribution of the random access code. Differently from [19], this work considers a channel code at the physical layer, which exists in most practical systems, and is shown to greatly influence the design and performance of the random access code.

II. System Model

In this work we denote scalars in lower case, vectors in bold lower case and matrices in bold upper case. A superscript 'T' denotes the transpose, a superscript '$*$' denotes the complex conjugate and a superscript 'H' denotes the conjugate transpose.

We consider a random access system consisting of a single base station with M antennas and K users, each one with a single antenna, see Fig. 1. Communication is performed by using slotted time, where each time slot consists of an uplink pilot phase, an uplink data phase and a downlink data phase, see Fig. 2. In each time slot, each user is active with probability p_a. There are τ orthogonal pilot sequences $\{\boldsymbol{s}\}$, each consisting of τ symbols $\boldsymbol{s} = [s(1)\ s(2) \dots s(\tau)]$. An active user selects a pilot sequence randomly from the τ available pilot sequences. Note that multiple users may choose the same pilot sequence. See Fig. 3 for an example of a random pilot schedule with $\tau = 2$ and $K = 3$. The channel between the k'th user and the BS during the n'th time slot is denoted $\boldsymbol{h}_{n,k} = [h_{n,k}(1)\ h_{n,k}(2) \dots h_{n,k}(M)]^T$, where $h_{n,k}\ \forall\, n, k$ are i.i.d. We consider systems that apply ideal power control, such

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TWC.2018.2873400, IEEE Transactions on Wireless Communications

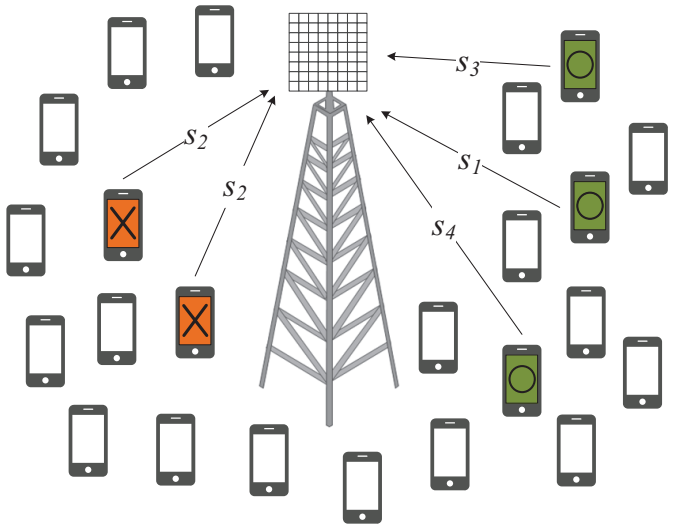

Fig. 1. A single cell crowd scenario. Red devices (crosses) experience interference due to colliding pilot signals. Green devices (circles) apply unique pilot sequences, whereby interference is avoided.

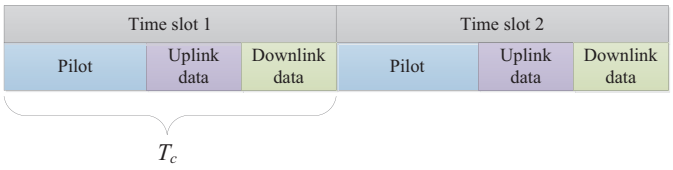

Fig. 2. An example of a transmission schedule.

that $h_{n,k}(i) \sim \mathcal{CN}(0,1)$, $\forall\, i$ [1]. The time slot has a duration T_c, which corresponds to the coherence time in which the channel coefficient remains constant. Let $\mathcal{A}_n$ denote all active users in time slot n, while $\mathcal{A}_n^j$ denotes the set of users that have selected $\boldsymbol{s}_j$ in the n'th time slot. If $\boldsymbol{Y}_n^{pu} \in \mathcal{C}^{M\times\tau}$, denotes the uplink pilot signal received in time slot n, we have

$$\boldsymbol{Y}_n^{pu} = \sum_{j=1}^{\tau} \sum_{k\in\mathcal{A}_n^j} \boldsymbol{h}_{n,k}\boldsymbol{s}_j + \boldsymbol{Z}_n^{pu}, \tag{1}$$

where $\boldsymbol{Z}_n^{pu}$ is a matrix of i.i.d. Gaussian noise components, hence $\boldsymbol{Z}_n^{pu}(i,j) \sim \mathcal{CN}(0,\sigma_n^2)$, $\forall\, i,j$. Any future instances of a vector $\boldsymbol{z}$ or matrix $\boldsymbol{Z}$, with different sub- or superscripts follow the same definition. All active users transmit a message of length D_u symbols in the uplink data phase. The message from the k'th user is denoted $\boldsymbol{x}_k^u = [x_k^u(1)\ x_k^u(2)\dots x_k^u(D_u)]$. Using $\boldsymbol{Y}_n^u \in \mathcal{C}^{M\times D_u}$, to denote the data part of the received signal in the uplink, we get:

$$\boldsymbol{Y}_n^u = \sum_{k\in\mathcal{A}_n} \boldsymbol{h}_{n,k}\boldsymbol{x}_k^u + \boldsymbol{Z}_n^u. \tag{2}$$

In the downlink phase we rely on channel reciprocity, such that the uplink channel estimate is assumed to be a valid estimate for the downlink transmission. The BS transmits a precoded

[1]Note that in a crowd scenario, independent channels across users may not be a valid assumption. However, in this work this assumption is only applied when processing signals from a small subset of the active crowd, in which case independent channels is a valid assumption.

	Time slot 1			Time slot 2			Time slot 3			Time slot 4			Time slot 5	
User 1	s_1	x_1^u		s_2	x_1^u		s_2	x_1^u					s_1	x_1^u
User 2	s_1	x_2^u		s_1	x_2^u	x_2^d	s_2	x_2^u						
User 3	s_2	x_3^u	x_3^d	s_2	x_3^u					s_1	x_3^u	x_3^d	s_1	x_3^u

Fig. 3. An example of a pilot schedule. Subscripts for pilots refer to indexing within the set of pilots, whereas subscripts for data refer to users.

downlink pilot symbol, such that the k-th user receives a downlink pilot signal, $y_{n,k}^{pd}$, given by

$$y_{n,k}^{pd} = \boldsymbol{h}_{n,k}^T\boldsymbol{w}_{n,k} + \boldsymbol{z}_{n,k}^{pd}, \tag{3}$$

where $\boldsymbol{w}_{n,k} = [w_{n,k}(1)\ w_{n,k}(2)\dots w_{n,k}(M)]^T = \boldsymbol{h}_{n,k}^*$ is the precoding vector for user k in the n'th time slot. Clearly, this assumes that the BS has an estimate from the uplink pilot of the channel $\boldsymbol{h}_{n,k}$ before the downlink transmission. We denote the set of users, for which the BS has an estimate of $\boldsymbol{h}_{n,k}$, as $\mathcal{C}_n$. Note that a single-symbol downlink pilot is sufficient, thus the associated rate loss can be neglected. We denote the downlink message intended for user k, $\boldsymbol{x}_k^d = \left[x_k^d(1)\ x_k^d(2)\dots x_k^d(D_d)\right]$ and define $\boldsymbol{X}^d$ with the k'th row given by $\boldsymbol{x}_k^d$ for $k \in \mathcal{C}_n$. Similarly we define $\boldsymbol{W}_n$ with k'th column being $\boldsymbol{w}_{n,k}$ and $\boldsymbol{H}_n$ with $\boldsymbol{h}_{n,k}$ as the k'th column for $k \in \mathcal{C}_n$. The received downlink data signal is then expressed as

$$\boldsymbol{Y}_n^d = \boldsymbol{H}_n^T\boldsymbol{W}_n\boldsymbol{X}^d + \boldsymbol{Z}_n^d. \tag{4}$$

In both uplink and downlink, the coherence time allows the transmission of L symbols and we have $L = \tau + D = \tau + D_u + D_d$. D_u and D_d can in principle be chosen arbitrarily for asymmetric operation as long as $D = L - \tau$, however in the remainder of this work we consider the case $D_u = D_d = \frac{D}{2}$. The data is assumed to be channel coded with rate R at the physical layer, such that the effective data rate is $R\frac{D}{2L}$. We consider arbitrary channel codes with hard detection decoding and therefore apply the upper bound on the error correction capabilities of such codes. More specifically, we consider a data message successfully recovered if $p_e \leq \frac{(1-R)}{2}$, where p_e is the bit error rate. Hence, the numerical results serve as upper bounds from a channel code perspective. By $\mathcal{S}_n$ we denote the set of users, whose associated data message, uplink or downlink, is successfully recovered. Note that $|\mathcal{S}_n| \leq \tau$ and $|\mathcal{S}_n| \leq |\mathcal{A}_n|$. The throughput of the system in time slot n, γ_n, is then defined as the sum-rate given by

$$\gamma_n = \frac{|\mathcal{S}_n|R(L-\tau)}{2L} = \frac{|\mathcal{S}_n|RD}{2\,L}. \tag{5}$$

Note that γ_n is defined as successfully recovered messages in a time slot. The modulation rate can be chosen arbitrarily and will influence the probability distribution of $|\mathcal{S}_n|$ and hence the throughput. Details on this are given in Section IV.

III. Coded Pilot Access

This section describes the proposed method of communication in the system described in Section II, treating both uplink and downlink operation.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TWC.2018.2873400, IEEE Transactions on Wireless Communications

A. Uplink

In uplink operation, transmissions are organized in blocks of Δ consecutive time slots, referred to as a *frame*. If a user is active multiple times within a frame, the uplink data is retransmitted, similar to conventional coded random access schemes. We introduce a parameter called the *overhead factor* α, defined as

$$\alpha = \frac{\tau\Delta}{K}, \tag{6}$$

which is an expression of the normalized amount of orthogonal resources in a frame. Ideally $\alpha = 1$, in which case there is exactly one orthogonal resource per user. However, at finite frame length, the scheme requires a small overhead in order to operate well. Hence, in practice α will attain values slightly above one. Details on this will follow in Section IV. The performance parameter of interest is the frame average uplink throughput given by $\gamma_u = \sum_{n=1}^{\Delta} \gamma_n/\Delta$.

From the uplink pilot signals in (1), it is possible to estimate the channels between the users and the base station. However, since multiple users may apply the same pilot sequence, it is only possible to estimate a sum of the involved channels. The least squares estimate, $\boldsymbol{\phi}_{n,j}$, based on the pilot signal in time slot n from users applying $\boldsymbol{s}_j$ is found as

$$\begin{aligned}\boldsymbol{\phi}_{n,j} &= (\boldsymbol{s}_j \boldsymbol{s}_j^H)^{-1} \boldsymbol{Y}_n^{pu} \boldsymbol{s}_j^H \\ &= \sum_{k \in \mathcal{A}_n^j} \boldsymbol{h}_{n,k} + \boldsymbol{z}_n^{pu'}.\end{aligned} \tag{7}$$

where $\boldsymbol{z}_n^{pu'}$ is the post-processed noise terms originating from $\boldsymbol{Z}_n^{pu}$. Any future instances of a vector $\boldsymbol{z}$ with a prime follow the same definition.

The problem of interfering users applying the same, or a non-orthogonal, pilot sequence is often called *pilot contamination*. If we proceed to detect the data in the uplink phase using a contaminated channel estimate, the result will be a summation of data messages. By $\boldsymbol{\psi}_{n,j}$ we denote the data estimate based on the channel estimate $\boldsymbol{\phi}_{n,j}$. Assuming orthogonality between user channels, i.e. $\lim_{M\to\infty} \boldsymbol{h}_{n,m}^H \boldsymbol{h}_{n,k}/M = 0$ almost surely for $m \neq k$ [27], we then have

$$\begin{aligned}\boldsymbol{\psi}_{n,j} &= (\boldsymbol{\phi}_{n,j}^H \boldsymbol{\phi}_{n,j})^{-1} \boldsymbol{\phi}_{n,j}^H \boldsymbol{Y}_n^u \\ &= \sum_{k \in \mathcal{A}_n^j} \frac{\boldsymbol{\phi}_{n,j}^H \boldsymbol{h}_{n,k}}{||\boldsymbol{\phi}_{n,j}||^2} \boldsymbol{x}_k^u + \boldsymbol{z}_n^{u'}.\end{aligned} \tag{8}$$

Hence, a pilot collision leads to a data collision, i.e. interference among data signals. A classical way to deal with this problem is to minimize the probability of contamination by carefully selecting p_a. The objective of such criterion is to maximize the probability of having only one user applying a particular pilot sequence in a particular time slot. Hence, we have

$$\begin{aligned} \underset{p_a}{\text{maximize}} \quad & \Pr(\left|\mathcal{A}_n^j\right| = 1) \\ \text{subject to} \quad & 0 \leq p_a \leq 1 \end{aligned} \tag{9}$$

This will maximize the number of non-contaminated channel estimates, and in turn maximize the number of successful data transmissions. This approach is reminiscent of the framed slotted ALOHA protocol for conventional random access. We consider this a reference scheme in this work and refer to it as ALOHA. Note that a random access, i.e. a nonscheduled scheme, must be considered as a reference, due to the assumption of a crowd scenario, where scheduling is infeasible.

A novel alternative solution is presented in this paper, which does not consider data collisions as waste, but instead buffers the collided signals and use them subsequently through an iterative process, whereby they contribute to the throughput. We call it *Coded Pilot Access* (CPA). This solution is based on applying the contaminated estimates as *matched filters* on the received uplink data signals, $\boldsymbol{Y}_n^u$. Denoting the filtered data signal $\boldsymbol{f}_{n,j} \in \mathcal{C}^{1\times D_u}$, we have

$$\begin{aligned}\boldsymbol{f}_{n,j} &= \boldsymbol{\phi}_{n,j}^H \boldsymbol{Y}_n^u \\ &= \sum_{k \in \mathcal{A}_n^j} \left(||\boldsymbol{h}_{n,k}||^2 + \sum_{m \in \mathcal{A}_n^j \setminus \{k\}} \boldsymbol{h}_{n,m}^H \boldsymbol{h}_{n,k} \right) \boldsymbol{x}_k^u \\ &\quad + \sum_{\ell \in \mathcal{A}_n \setminus \mathcal{A}_n^j} \left(\sum_{o \in \mathcal{A}_n^j} \boldsymbol{h}_{n,o}^H \boldsymbol{h}_{n,\ell} \right) \boldsymbol{x}_\ell^u + \boldsymbol{z}_n^{u'}.\end{aligned} \tag{10}$$

Note that compared to the detection in (8), we do not normalize with the power of the channel estimate. In the case of no collision, normalization will provide the data message directly. However, in the case of a collision, the filtered data signal involves multiple data messages and the channel power involves multiple user channels, whereby normalization is futile. Instead we end detection after the matched filtering step and buffer the signals. By relying on two essential features from the massive MIMO scenario, (10) can be simplified greatly, when M goes towards infinity. The first feature is orthogonality between user channel vectors. This implies that $\lim_{M\to\infty} \boldsymbol{h}_{n,m}^H \boldsymbol{h}_{n,k}/M = 0$ almost surely for $m \neq k$. The second feature is the temporal stability of channel powers, which implies that $||\boldsymbol{h}_{n,k}||^2 = ||\boldsymbol{h}_{n',k}||^2 \ \forall\ n, n'$. This allows us to drop the time index in the channel powers. We thus have the following expression for the filtered data signal in the limit of $M \to \infty$:

$$\lim_{M\to\infty} \boldsymbol{f}_{n,j} = \sum_{k \in \mathcal{A}_n^j} ||\boldsymbol{h}_k||^2 \boldsymbol{x}_k^u + \boldsymbol{z}_n^{u'}. \tag{11}$$

Hence, the implications of pilot contamination has been turned into linear combinations of data messages, through post-processing with matched filters. The coefficients of the linear combinations are the temporally stable channel powers. By again relying on the asymptotic properties of the massive MIMO channel, the channel vector estimates in (7) can be utilized to find estimates of the sums of the channel powers. We denote these as $g_{n,j}$ and have

$$\begin{aligned} g_{n,j} &= \boldsymbol{\phi}_{n,j}^H \boldsymbol{\phi}_{n,j} \\ &= \sum_{k \in \mathcal{A}_n^j} \left(||\boldsymbol{h}_{n,k}||^2 + \sum_{m \in \mathcal{A}_n^j \setminus \{k\}} \boldsymbol{h}_{n,m}^H \boldsymbol{h}_{n,k} \right) + \boldsymbol{z}_n^{pu'}, \\ \lim_{M\to\infty} g_{n,j} &= \sum_{k \in \mathcal{A}_n^j} ||\boldsymbol{h}_k||^2 + \boldsymbol{z}_n^{pu'}. \end{aligned} \tag{12}$$

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TWC.2018.2873400, IEEE Transactions on Wireless Communications

Eqs. (11) and (12) for $n = 1, \ldots, \Delta$, and $j = 1, \ldots, \tau$, represent a system of equations, which we wish to solve for $\boldsymbol{x}_k^u$, $k = 1, \ldots, K$. It should be noted that the BS has no a-priori knowledge of the random activity and pilot choices of the users. Hence, the system of equations cannot be solved using, e.g. Gaussian elimination. Instead we employ *successive interference cancellation* (SIC), as in recent works on CRA [28].

SIC proceeds as follows. Initially, the BS locates immediately decodable uplink data[2], i.e. cases of $\left|\mathcal{A}_n^j\right| = 1$. If $\boldsymbol{f}_{n,j}$ is decodable, we furthermore have an estimate of the channel norm of the transmitter in $g_{n,j}$. We also assume that the uplink data embeds the information about the random pilot and data transmission schedule of the transmitter[3], which allows the BS to locate all the replicas of the same packet sent by that transmitter. Hence, although the BS has no a-priori knowledge of the pilot choices, it is available before interference cancellation. In the context of (11) and (12), when the data from user k is successfully decoded, the BS learns for which n and j we have $k \in \mathcal{A}_n^j$. This enables the BS to cancel the interference caused by the replicas from user k by subtracting $||\boldsymbol{h}_k||^2 \boldsymbol{x}_k^u$ from any $\boldsymbol{f}_{n,j}$ for which $k \in \mathcal{A}_n^j$. Furthermore, the interference caused by the associated pilot transmissions can be canceled by subtracting $||\boldsymbol{h}_k||^2$ from any $g_{n,j}$ for which $k \in \mathcal{A}_n^j$. The cancellations cause k to be removed from any $\mathcal{A}_n^j$ it originally appeared in. Potentially, this leads to new cases of $\left|\mathcal{A}_n^j\right| = 1$, whereby new data can be recovered, and the iterative process can continue. The uplink operation is described in Algorithm 1.

The employed decoding algorithm is analogous to belief propagation (BP) decoding of erasure codes. A common way of visualizing such codes is by using bipartite graphs. They also apply in our context, see Fig. 4 for an example based on the first two time slots in the example from Fig. 3. Squares are referred to as factor-nodes and represent observable signals after matched filtering. Hence, each factor-node corresponds to an orthogonal resource, i.e. a pilot in a time slot within a frame. Circles are referred to as variable-nodes and represent data messages, which we wish to recover. An edge connecting a variable-node with a factor-node represents a transmission of the replica of the data message, and denotes that the variable is a part of the linear combination represented by the factor-node. The number of edges connected to a node is referred to as the degree of the node. Based on Fig. 4, we can walk through the simple example of decoding $\boldsymbol{x}_1^u$.

Example: Initially the BS detects that $\boldsymbol{f}_{12}$ has degree one, and thereby directly recovers $\boldsymbol{x}_3^u$. The data from user 3 makes the BS aware of the activity pattern of this user and thereby enables cancellation of its interference. As a result, $||\boldsymbol{h}_3||^2$ is subtracted from g_{22} and $||\boldsymbol{h}_3||^2 \boldsymbol{x}_3^u$ is subtracted from $\boldsymbol{f}_{22}$. This cancellation has reduced the degree of $\boldsymbol{f}_{22}$ to one, which is detected by the BS, whereby $\boldsymbol{x}_1^u$ is recovered.

In the example, noise is assumed to not garble the decoding of the rate R channel code applied at the physical layer. In other words, when a signal has been reduced to degree one, the corresponding data is recovered successfully. Clearly, this is a strong assumption that cannot always hold. In fact, signals of higher degree have a higher risk of being undecodable at the physical layer after being reduced to degree one. The reason is accumulation of noise during interference cancellation. While interference cancellation greatly increases the SINR, it actually

[2]In practice this is enabled by applying a cyclic redundancy check (CRC) code on the data.

[3]A practical solution to this is to embed the seed for the random number generator. The rate loss due to the CRC and the embedded seed is considered negligible.

Algorithm 1 Coded Pilot Access (Uplink)

User Equipment:

for $n = 1$ to Δ **do**
  $r \leftarrow unif(0, 1)$ (uniform random variable between 0 and 1).
  if $r \leq p_a$ **then**
    Select at random one of τ pilot sequences.
    Transmit uplink pilot followed by data message.
  end if
end for

Base Station:

Buffer $\leftarrow \emptyset$
for $n = 1$ to Δ **do**
  Receive $\boldsymbol{Y}_n^{pu}$ and $\boldsymbol{Y}_n^u$.
  for $j = 1$ to τ **do**
    $\boldsymbol{\phi}_{n,j} \leftarrow (\boldsymbol{s}_j \boldsymbol{s}_j^H)^{-1} \boldsymbol{Y}_n^{pu} \boldsymbol{s}_j^H$
    $\boldsymbol{f}_{n,j} \leftarrow \boldsymbol{\phi}_{n,j}^H \boldsymbol{Y}_n^u$
    $g_{n,j} \leftarrow \boldsymbol{\phi}_{n,j}^H \boldsymbol{\phi}_{n,j}$
    if $\frac{\boldsymbol{f}_{n,j}}{g_{n,j}}$ is decodable at physical layer **then**
      Add $\boldsymbol{x}_k^u$ and $||\hat{\boldsymbol{h}}_k||^2 = g_{n,j}$ to Buffer.
    end if
  end for
end for
while Buffer $\neq \emptyset$ **do**
  Access buffered data, $\boldsymbol{x}_k^u$.
  Access buffered channel norm estimate, $||\hat{\boldsymbol{h}}_k||^2$.
  Extract pilot activity pattern of user k embedded in $\boldsymbol{x}_k^u$.
  for $n = 1$ to Δ **do**
    for $j = 1$ to τ **do**
      if $k \in \mathcal{A}_n^j$ **then**
        $\boldsymbol{f}_{n,j} \leftarrow \boldsymbol{f}_{n,j} - ||\hat{\boldsymbol{h}}_k||^2 \boldsymbol{x}_k^u$
        $g_{n,j} \leftarrow g_{n,j} - ||\hat{\boldsymbol{h}}_k||^2$
        if $\frac{\boldsymbol{f}_{n,j}}{g_{n,j}}$ is decodable at physical layer **then**
          Add $\boldsymbol{x}_\ell^u$ and $||\hat{\boldsymbol{h}}_\ell||^2 = g_{n,j}$ to Buffer.
        end if
      end if
    end for
  end for
end while

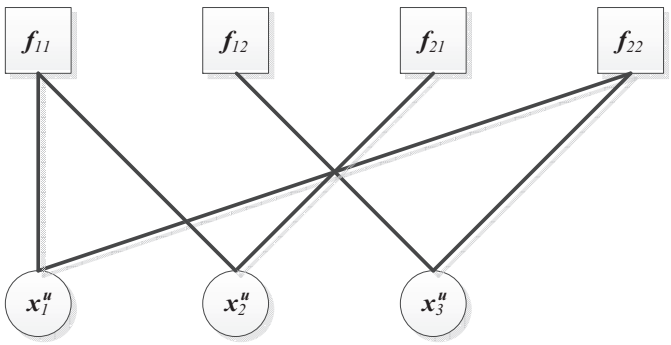

Fig. 4. A bipartite graph representation of the data collisions.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TWC.2018.2873400, IEEE Transactions on Wireless Communications

decreases the SNR, which potentially makes a rate R channel code undecodable. This effect is analyzed in section IV.

The performance of the BP decoder for erasure codes is tightly connected with the factor- and variable-node *degree distribution*, denoted as Ψ and Λ, respectively, where Ψ_d/Λ_d is the probability that a factor/variable-node has degree d. Several works [29], [30] have studied the design of well performing degree distributions. However, in this context we do not have the full freedom to tailor the degree distributions. Our only way of influencing degree distributions is through the choice of p_a and the overhead factor α, see (6). Specifically, since a user is applying a particular pilot sequence in a particular time slot with probability p_a/τ, and there are Δ time slots, we have the following relation between the degree distributions and p_a and α:

$$\Psi_d = \Pr(\left|\mathcal{A}_n^j\right| = d) = \binom{K}{d}\left(\frac{p_a}{\tau}\right)^d\left(1-\frac{p_a}{\tau}\right)^{K-d} \quad (13)$$

$$\approx \frac{(\frac{p_a K}{\tau})^d}{d!}e^{-\frac{p_a K}{\tau}} = \frac{\beta^d}{d!}e^{-\beta}, \quad (14)$$

where β is the average factor-node degree

$$\beta = \frac{p_a K}{\tau} \quad (15)$$

and

$$\Lambda_d = \binom{\Delta}{d}p_a^d(1-p_a)^{\Delta-d} \approx \frac{(\Delta\, p_a)^d}{d!}e^{-\Delta p_a} \quad (16)$$

$$= \frac{(\alpha\beta)^d}{d!}e^{-\alpha\beta}, \quad (17)$$

where

$$\alpha = \frac{\tau\Delta}{K}. \quad (18)$$

Obviously, through the choice of β (i.e. the choice of p_a) and α (i.e. the choice of Δ), one determines the degree distributions. In section IV we provide the analytical optimization of β and α.

B. Downlink

In order to choose an appropriate precoder for the downlink transmission, the BS must have an estimate of the current channel. The coded operation applied in uplink, which results in multiple collisions and occasional single (non-contaminated) transmissions, does not guarantee that such an estimate is available. Uplink operation relies on SIC based only on knowledge of the norm. Hence, downlink transmission to a user is only possible if that user avoided collision during the previous uplink pilot phase, such that an uncontaminated channel estimate is available.

Regarding the reception of a downlink transmission and assuming channel reciprocity, the k'th user does not need to estimate each coefficient of $\boldsymbol{h}_{n,k}$, which would require a pilot signal for all M antennas. Instead, we let the receiver estimate the concatenated "channel" consisting of both the downlink precoder, $\boldsymbol{W}_n$, and the actual channel of user k, $\boldsymbol{h}_{n,k}$. Denoting the concatenated channel, $\boldsymbol{q}_{n,k}$, we have

$$\boldsymbol{q}_{n,k} = \boldsymbol{h}_{n,k}^T\boldsymbol{W}_n, \quad (19)$$

where $\boldsymbol{q}_{n,k}$ is estimated through (3). The part of the downlink signal, $\boldsymbol{Y}_n^d$, received by user k is denoted $\boldsymbol{y}_{n,k}^d$. We have

$$\begin{aligned}\boldsymbol{y}_{n,k}^d &= \boldsymbol{h}_{n,k}^T\boldsymbol{W}_n\boldsymbol{X}^d + \boldsymbol{z}_{n,k}^d,\\ &= \boldsymbol{q}_{n,k}\boldsymbol{X}^d + \boldsymbol{z}_{n,k}^d. \end{aligned} \quad (20)$$

Hence, utilizing the estimate of $\boldsymbol{q}_{n,k}$, the k'th user is able to recover its part, $\boldsymbol{x}_k^d$ of the transmitted message $\boldsymbol{X}^d$, and subsequently attempt decoding at the physical layer.

IV. Analysis

This section presents an analysis of the throughput and latency of the proposed scheme, CPA, and the reference scheme, ALOHA.

A. CPA

The SIC algorithm described in Section III can be analyzed using the state-of-the-art analytical tools devised for BP erasure decoding, specifically, using the and-or tree evaluation [31].[4] For the given factor node degree distributions Ψ and variable node degree distribution Λ, the and-or tree evaluation outputs the asymptotic probability, when $K \to \infty$ and $\Delta \propto K$, of recovering a data message (i.e., user signal), in a scenario when the impact of noise can be neglected and removal of previously recovered messages from factor nodes is perfect. However, in the scenario assessed in the paper there are important differences that have to be taken into account, which stem from the nature of the physical layer operation:

- The decodability of user signals received in singleton slots depends on the received SNR.
- The cancellation of decoded signals is not ideal and leaves residual interference power. This implies that, as the SIC progresses, the accumulated residual interference effectively decreases SNR, which may prevent decoding of the signals whose degree become reduced to one.

The and-or tree evaluation assumes that the bipartite graph representation, an example of which is given in Fig. 4, can be unfolded into a tree, by taking a randomly chosen variable-node as the root of the tree and successively adding into the tree adjacent factor-nodes and then adjacent variable-nodes using a breadth-first search method. Fig. 5 provides a generic example of such a tree. The reception algorithm is modelled as the iterative application of the following two operations, starting from the leaves of the tree:

(i) decoding of edges of emanating from factor-nodes (i.e., decoding of replicas of user signals) and thus, decoding the related variable-nodes to which the edges are incident, corresponding to (a generalized) "and" operation, cf. [20], [33],

(ii) removal of edges emanating from decoded variable-nodes (i.e., removal of other replicas of decoded signals) from the factor nodes in the next level of the tree, corresponding to the "or" operation.

[4]For a general introduction to the and-or tree evaluation, we refer the interested reader to [31], [32].

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TWC.2018.2873400, IEEE Transactions on Wireless Communications

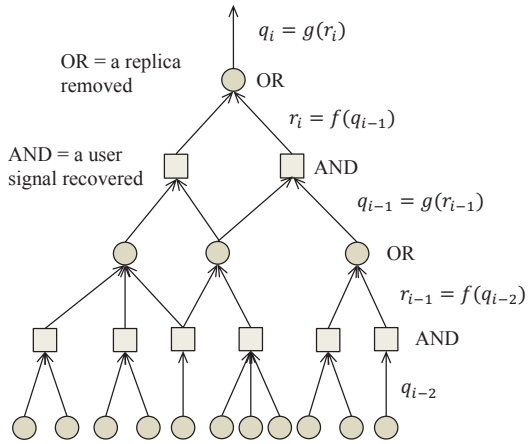

Fig. 5. A tree representation of the iterative reception algorithm.

Both operations are probabilistically characterized, in terms of expected probability of *not* decoding a user signal in a factor-node in i'th iteration, denoted as r_i, and *not* removing a replica in i'th iteration, denoted as q_i, respectively

$$r_i = \sum_j \psi_j \, r_{i|j}, \tag{21}$$

$$q_i = \sum_k \lambda_k \, q_{i|k} \tag{22}$$

where ψ_j is the probability that an edge is connected to a factor-node of degree j, $r_{i|j}$ is the probability that the replica of the user signal represented by an edge is not decoded in the factor-node of degree j, λ_k is the probability that an edge is connected to a variable-node of degree k, and $q_{i|k}$ is the probability that an edge (i.e., replica) emanating from a variable-node of degree k is not removed. The tree structure allows for the successive updates of these expected probabilities, as depicted in Fig. 5. Here we remark that the formal proof of the correctness of this approach that is based on expected probabilities (i.e., assuming that the all edges are statistically equal, independent of the actual degrees of the nodes that they are connected to) for the and-or tree evaluation can be found in [32]. We also note that in the non-asymptotic case, the graph representation contains loops, and the corresponding tree representation is only an approximation, where the obtained results present an upper bound on the non-asymptotic performance, see [32].

We continue by introducing the edge-oriented degree distributions [31], corresponding to probabilities that a randomly chosen edge in the graph is connected to a node of a certain degree. In particular, these are the edge-oriented factor-node degree distribution ψ and the edge-oriented variable-node degree distribution λ, which can be derived through the factor- and variable-node degree distributions Ψ and Λ, respectively [31]

$$\psi_d = \frac{d \, \Psi_d}{\sum_j j \, \Psi_j}, \; d \geq 1, \tag{23}$$

$$\lambda_d = \frac{d \, \Lambda_d}{\sum_j j \, \Lambda_j}, \; d \geq 1. \tag{24}$$

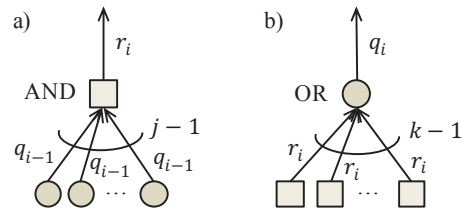

Fig. 6. Probability updates in a) factor node and b) variable node.

Assume a factor node of degree j. The probability that an edge connected to a factor node of degree j is not removed in the i'th iteration, denoted by $r_{i|j}$, is

$$r_{i|j} = \pi_j \, (1 - q_{i-1})^{j-1}, \; i \geq 1, \tag{25}$$

where π_j is the probability of recovering a user signal in the factor node of degree j when $j-1$ interfering signals were cancelled[5], and the term $(1-q_{i-1})^{j-1}$ refers to the probability that $j-1$ interfering signals were cancelled, see Fig. 6 a). The impact of the physical layer, i.e. receiver operation, as described in Section III, is embedded in π_j.[6] Averaging over the edge-oriented factor-node degree distribution yields

$$r_i = \sum_j \psi_j \, r_{i|j} = \sum_j \psi_j \, \pi_j \, (1 - q_{i-1})^{j-1}, \; i \geq 1. \tag{26}$$

Further, the probability that an edge connected to a variable node of degree k is not removed in the i'th iteration, denoted by $q_{i|k}$, is

$$q_{i|k} = r_i^{k-1}, \; i \geq 1, \tag{27}$$

where r_i^{k-1} refers to the probability that none of the $k-1$ replicas were recovered, see Fig. 6 b). Averaging over the edge-oriented variable-node degree distribution produces

$$q_i = \sum_k \lambda_k \, q_{i|k} = \sum_k \lambda_k \, r_i^{k-1}, \; i \geq 1, \tag{28}$$

with the initial value $q_0 = 1$. Combining (26) and (28), we get

$$q_i = \sum_k \lambda_k \Big(\sum_j \psi_j \, \pi_j \, (1 - q_{i-1})^{j-1} \Big), \; i \geq 1. \tag{29}$$

The output of the evaluation is the probability that a user signal becomes recovered:

$$p_d = 1 - \lim_{i \to \infty} q_i. \tag{30}$$

Obviously, (29) depends on the edge-oriented degree distributions ψ, λ and probabilities π_j. While the latter depends on the physical layer operation, ψ and λ can be optimized through optimization of Ψ and Λ, which in turn are optimized by optimizing β and α, see (15) and (6). Specifically, in the proposed scheme, we optimize α and β in order to maximize

[5]I.e. the probability of recovering a user signal from a factor node whose original degree j is reduced to 1.

[6]In the standard and-or tree evaluation, $\pi_j = 1$, $j \geq 1$, i.e. no matter what was the original degree of the factor node j, once it has been reduced to a singleton, the remaining signal is perfectly decoded. However, in the scenario assumed in the paper, this is not the case and π_j is a function of j.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TWC.2018.2873400, IEEE Transactions on Wireless Communications

the expected throughput γ_u, see Section III-A, which can be expressed as

$$\begin{aligned}\gamma_u &= \sum_{n=1}^{\Delta} \frac{\gamma_n}{\Delta} \\ &= \sum_{n=1}^{\Delta} \frac{|\mathcal{S}_n|}{\Delta} R \frac{L-\tau}{L} \\ &= \frac{p_d K}{\Delta} R \frac{L-\tau}{L} \\ &= \frac{p_d R (L-\tau)\tau}{2\alpha L}. \end{aligned} \tag{31}$$

As described in section III-B, downlink transmission is only possible in a time/pilot resource if the associated uplink pilot transmission was uncontaminated. This occurs with probability Ψ_1, given in equation (14). We furthermore require that the uplink data transmission is successfully decoded at the physical layer, in order to identify the user. The probability of this is given by π_1, which is also the probability that the downlink transmission is successfully decoded (due to channel reciprocity). We can thus express the downlink throughput, γ_d, as follows:

$$\gamma_d = \frac{\Psi_1 \pi_1^2 R (L-\tau)\tau}{2L}. \tag{32}$$

The expected uplink latency, Ω_u, is given by the expected number of time slots necessary for a successful uplink transmission. We assume that the number of users, K, remains constant across frames and that users failing to transmit in uplink continue to attempt in the following frame. We furthermore assume that decoding is only attempted at the end of a frame, which limits decoding complexity and only has a minor impact on latency [28]. During a frame of $\Delta = \frac{\alpha K}{\tau}$ time slots, a user will successfully transmit an uplink message with probability p_d. The number of necessary frames thus follows the geometric distribution, with mean value $\frac{1}{p_d}$, which means Ω_u can be expressed as:

$$\Omega_u = \frac{\alpha K}{p_d \tau}. \tag{33}$$

The expected downlink latency, Ω_d, is similarly given by the expected number of time slots necessary for a successful downlink transmission. This requires an uncontaminated uplink pilot transmission, followed by a successful uplink data transmission and downlink transmission. This occurs with probability $p'_a = \pi_1^2 p_a \left(1 - \frac{p_a}{\tau}\right)^{K-1}$. The number of necessary time slots thus follows the geometric distribution, with mean value $\frac{1}{p'_a}$, which means Ω_d can be expressed as:

$$\Omega_d = \frac{1}{\pi_1^2 p_a \left(1 - \frac{p_a}{\tau}\right)^{K-1}}. \tag{34}$$

There is a natural tradeoff between optimizing p_a for high uplink throughput (low uplink latency) and optimizing it for high downlink throughput (low downlink latency). Such a joint optimization is outside the scope of this work, however in section V, we will provide numerical insight on the trade-off.

B. ALOHA

The optimization problem in (9) consists of maximizing the probability of a binomial random variable attaining the value one, see eq. (23). It can easily be shown that the solution to this optimization is $p_a = \frac{\tau}{K}$, such that the ALOHA uplink throughput is given by

$$\begin{aligned}\gamma_u^A &= \frac{\Psi_1 \pi_1 R (L-\tau)\tau}{L} \\ &= \left(1 - \frac{1}{K}\right)^{K-1} \pi_1 \frac{R(L-\tau)\tau}{2L}. \end{aligned} \tag{35}$$

The downlink throughput is then given by

$$\gamma_d^A = \left(1 - \frac{1}{K}\right)^{K-1} \pi_1^2 \frac{R(L-\tau)\tau}{2L}. \tag{36}$$

The uplink and downlink latencies follow the same derivation as eq. (34), such that since $p_a = \frac{\tau}{K}$, we have

$$\Omega_u^A = \frac{1}{\pi_1 \frac{\tau}{K} \left(1 - \frac{1}{K}\right)^{K-1}}, \tag{37}$$

$$\Omega_d^A = \frac{1}{\pi_1^2 \frac{\tau}{K} \left(1 - \frac{1}{K}\right)^{K-1}}. \tag{38}$$

We conclude by noting that the probabilities π_j, $j \geq 1$, are intractable to express analytically. Consequently, we evaluate π_j using Monte Carlo simulations for the analytical results in the following section.

V. Numerical Results

This section presents numerical results based on simulations and evaluations of the analysis in section IV. We compare the CPA scheme with the ALOHA random access scheme as described in connection with (9) and a scheduled conventional massive MIMO scheme, referred to as SMM. The SMM scheme is assumed to guarantee interference free transmissions, i.e. $\left|\mathcal{A}_n^j\right| = 1 \ \forall \ n, j$, thus not needing SIC. Users are assigned resources in a round-robin fashion. This scheme is considered an upper bound for a random access scheme.

The relevant parameters can be divided into two groups; system parameters and scheme parameters. System parameters are assumed to be given, whereas scheme parameters can be optimized for maximum throughput. We denote optimized parameters with a superscript $\star$, e.g. $\alpha^\star$ is the overhead factor, which maximizes the throughput. The throughput resulting from optimized parameters is denoted as $\gamma^\star$. All evaluations consider the case of $K = 1000$, QPSK modulation with power control and $\sigma_n^2 = 0.1$, i.e. an SNR of 10 dB.

System Parameters		Scheme Parameters	
σ_n^2	Noise power	τ	Pilot sequence length
K	Users in cell	α	Overhead factor
L	Coherence time	β	Avg. factor node degree
M	Antennas at BS	R	Channel code rate

Initially, we provide numerical insight to the trade-off between uplink and downlink performance. Fig. 7 shows throughput and latency of the proposed scheme for both uplink

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TWC.2018.2873400, IEEE Transactions on Wireless Communications

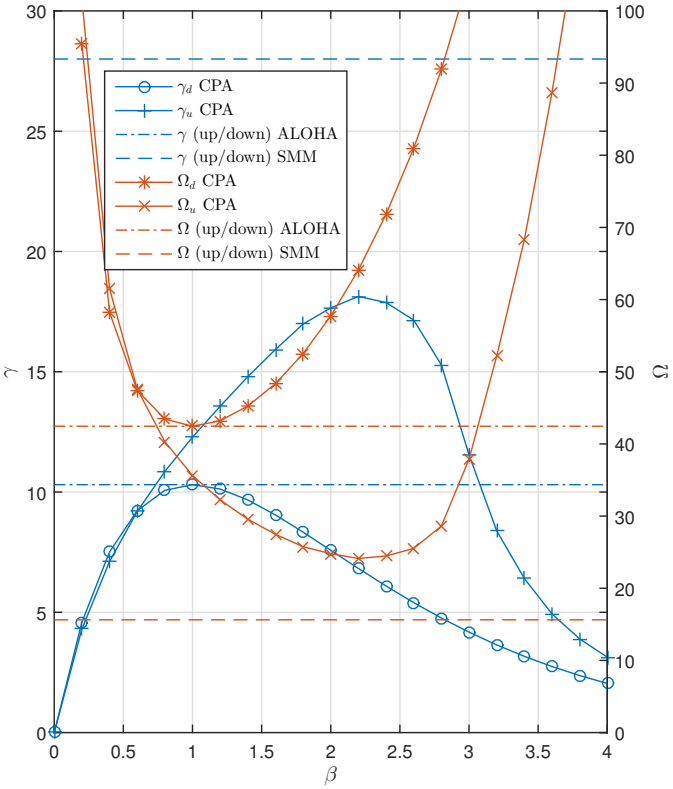

Fig. 7. Throughput (γ) and latency (Ω) as functions of β for both uplink and downlink. Fixed parameters are $\tau = 64$, $L = 512$, $M = 400$, $R = 0.5$ and $\alpha = 1.2$.

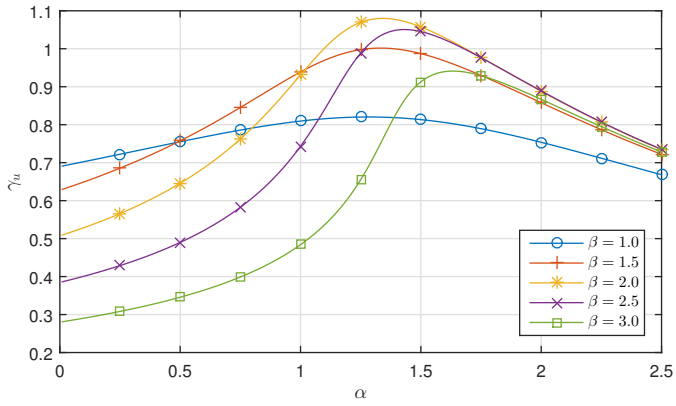

Fig. 8. And-or tree evaluation of throughput as a function of α for different values of β. Fixed parameters are $\tau = 4$, $L = 64$, $M = 400$ and $R = 1$.

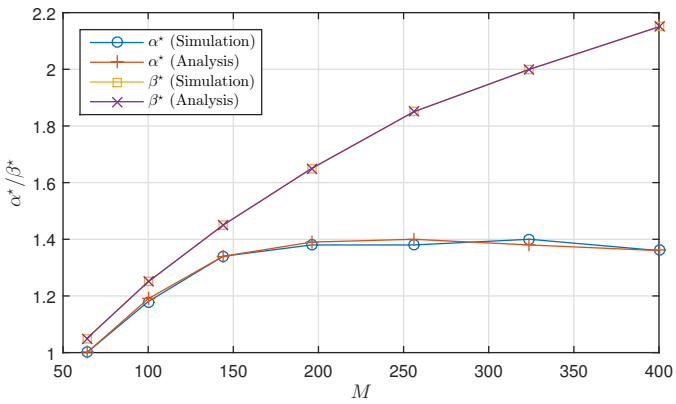

Fig. 9. Optimal values of α and β as a function of the number of antennas at the BS. Fixed parameters are $\tau = 4$, $L = 64$ and $R = 1$.

and downlink as a function of β, which is directly related to the activation probability, p_a, as expressed in eq. (15). Other parameters are fixed as $\tau = 64$, $L = 512$, $M = 400$, $R = 0.5$ and $\alpha = 1.2$. The throughput and latencies of the reference schemes are also included. SMM has an uplink and downlink throughput given by $\frac{R(L-\tau)\tau}{L}$ and an uplink and downlink latency given by $\frac{K}{\tau}$. We note that, in this case and most others, $\pi_1 = 1$, such that $\gamma_u^A = \gamma_d^A$ and $\Omega_u^A = \Omega_d^A$. The figure shows that the throughput and, especially, latency performance significantly depends on β. At $\beta = 1$, CPA matches the performance of ALOHA in downlink and slightly outperforms ALOHA in uplink. Increasing β provides a significant gain in uplink performance at a smaller expense in downlink performance. The remainder of this section focuses on the uplink throughput performance, since this is the relevant performance parameter for the novel aspects of our random access scheme to pilot sequences.

Next, we present results on numerical optimizations of the parameters α and β. Fig. 8 shows the uplink throughput, computed using and-or tree evaluation, as a function of α for different values of β, with $\tau = 4$, $L = 64$, $M = 400$ and $R = 1$, i.e. no physical layer channel code. It is clear that the value of α has a great impact on the performance of the CPA scheme. Performance peaks at α slightly above 1, after which throughput decreases. This is the point at which most user messages can be resolved with SIC, relative to the invested overhead α. Increasing α further, thus adding more resources to the frame, will just be a waste.

Fig. 9 shows the optimal values of α and β as a function of M, with $\tau = 4$, $L = 64$ and $R = 1$. The optimization is performed by evaluating the uplink throughput in ranges of parameter values, which, from existing work [20], [21], was expected to include and indeed included the global optimum. Both analytical and simulation results are included and shown to correspond very well. It is seen that increasing M allows for an increasing β, which indicates that SIC is better able to operate reliably. This is a result of improved orthogonality between user channels and improved temporal stability of the channel powers, which are essential properties as described in connection with equations (11) and (12).

Next, we turn our attention to the choice of a well performing value of τ. Note that α is proportional to the product of τ and Δ, see (5). Therefore, we now consider the optimal way to reach the desired α through the choice of τ and thereby Δ. On the one hand, increasing τ, while keeping α fixed, provides the same orthogonal resources using less time slots, which increases the potential throughput. On the other hand, increasing τ entails a rate loss, due to pilot symbols. Moreover, it entails a decrease in SINR, due to the increased number of orthogonal resources per time slot, which, for a given value of β, means an increased number of active users. This trade-off creates a correlation between the optimal τ and the value of M, as is seen in Fig. 10, where results for $L = 512$

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TWC.2018.2873400, IEEE Transactions on Wireless Communications

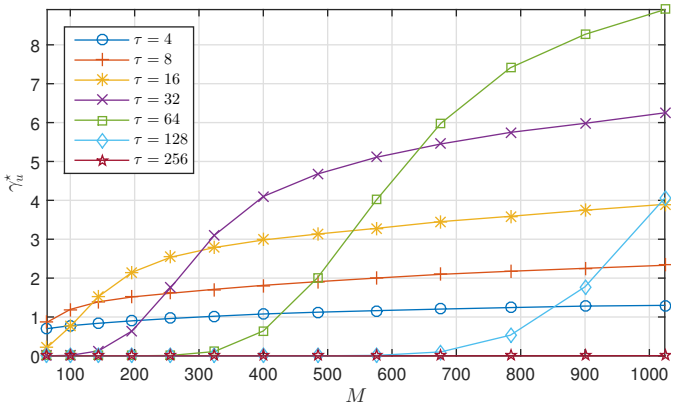

Fig. 10. Throughput of the proposed CPA scheme as a function of the number of antennas at the BS for different values of τ. Fixed parameters are $L = 512$, $R = 1$ and both α and β have been numerically optimized for each data point.

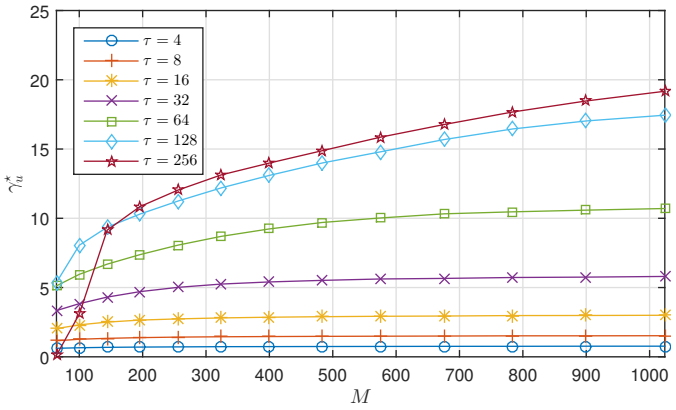

Fig. 11. Throughput of the proposed CPA scheme as a function of the number of antennas at the BS for different values of τ. Fixed parameters are $L = 512$, $R = 0.5$ and both α and β have been numerically optimized for each data point.

and $R = 1$ are plotted. Increasing M compensates for the decrease in SINR, such that SIC can operate at a higher τ. The throughput gain from this is seen to be quite significant. Fig. 11 illustrates the case where a rate 0.5 channel code is applied at the physical layer. In this case, the scheme can cope with lower SINR levels, and thus higher values of τ. Obviously, the drawback is the rate loss from the channel code. However, the gain of SIC overcompensates the rate loss from the channel code, whereby significantly higher uplink throughput is achieved compared to uncoded operation.

Finally, we present a comparison between the proposed CPA scheme and the two references, SMM and ALOHA. For all schemes, τ, α and β have been optimized. Fig. 12 shows results for both $R = 0.5$ and $R = 1$ with $L = 512$. As expected, the performance of all schemes increases with M. However, in the case of $R = 0.5$, the ALOHA scheme experiences a saturation of the performance at roughly $M = 200$, whereas the CPA scheme continues to increase. The reason is that the ALOHA scheme can only benefit from the increased SINR until the point, where degree one signals are decoded with high probability. The CPA is able to further benefit, due to improved SIC. Roughly a doubling of the uplink throughput is achieved

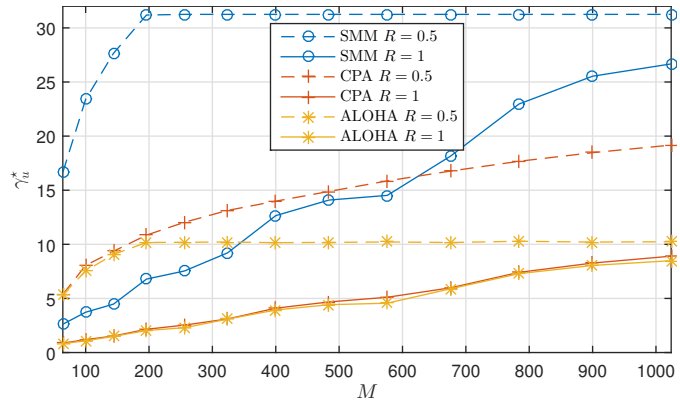

Fig. 12. Comparison of the evaluated schemes at $R = 0.5$ and $R = 1$ and optimized values of τ, α and β.

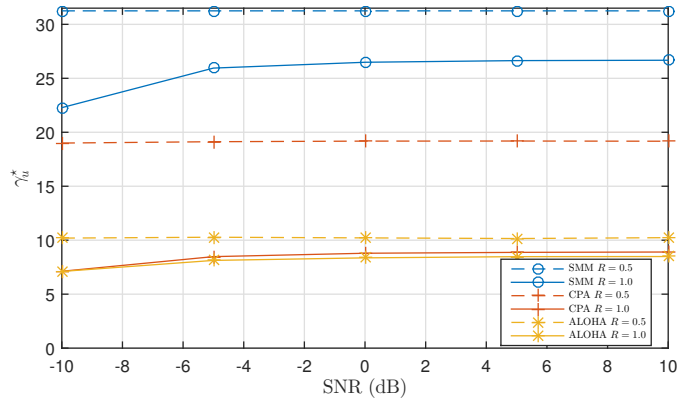

Fig. 13. Comparison of the evaluated schemes at $M = 1024$ for a range of SNR values.

at $M = 1024$ and $R = 0.5$ compared to ALOHA, which closes a significant part of the gap to the upper bound given by scheduled operation. For $R = 1$, a similar saturation of the performance of ALOHA is expected at higher M, outside the simulated interval. Since no channel code is applied, it requires much more antennas for degree one signals to be highly reliable.

Fig. 13 shows the same comparison but for a range of SNR values and $M = 1024$. This shows that the system, regardless of access scheme, is robust towards variations in the SNR. Only for uncoded operation, the system experiences a slight decrease of throughput at low SNR. Fig. 14 shows results for the same evaluations, but in a multi-cell scenario with a cell of interest surrounded by six neighboring cells in a hexagonal grid. All cells apply the same set of pilots. Each cell has a radius to a vertex of 1600 meters and a path loss exponent of 3.8. 1000 users are distributed uniformly at random in each cell, such that not only the cell of interest is crowded, but also the neighboring cells. The results show that CPA is still able to outperform ALOHA in this scenario, which is particularly challenging for our SIC based scheme. For $R = 0.5$, CPA outperforms ALOHA by 30 % in the simulated SNR range.

VI. Conclusions

Crowd scenarios present a particularly challenging access problem in massive MIMO systems. The intermittent traffic

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TWC.2018.2873400, IEEE Transactions on Wireless Communications

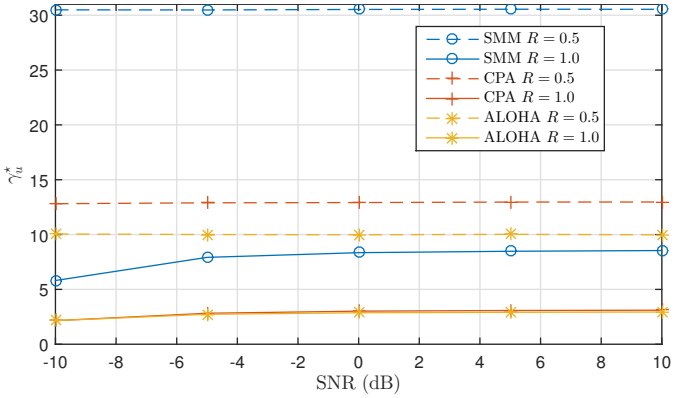

Fig. 14. Comparison of the evaluated schemes at $M = 1024$ for a range of SNR values in a multi-cell scenario.

from users and the scarcity of pilot sequences makes orthogonal scheduling infeasible. We presented a solution based on coded random access to the pilot sequences, which leverages on the channel hardening properties of massive MIMO. These allow us to view a set of contaminated pilot signals as a graph code on which iterative belief propagation can be performed. Using the tool of the and-or tree evaluation, we were able to analytically optimize the degree distribution of the random access code. With optimized parameters, the proposed solution proves highly efficient, comfortably outperforming the conventional ALOHA approach to random pilot access.

REFERENCES

[1] F. Boccardi, R. Heath, A. Lozano, T. Marzetta, and P. Popovski, "Five disruptive technology directions for 5G," *Communications Magazine, IEEE*, vol. 52, pp. 74–80, February 2014.

[2] T. Marzetta, "How much training is required for multiuser MIMO?," in *Signals, Systems and Computers, 2006. ACSSC '06. Fortieth Asilomar Conference on*, pp. 359–363, Oct 2006.

[3] H. Q. Ngo and E. G. Larsson, "EVD-based channel estimation in multicell multiuser mimo systems with very large antenna arrays," in *2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 3249–3252, March 2012.

[4] R. R. Müller, L. Cottatellucci, and M. Vehkaper, "Blind pilot decontamination," *IEEE Journal of Selected Topics in Signal Processing*, vol. 8, pp. 773–786, Oct 2014.

[5] D. Hu, L. He, and X. Wang, "Semi-blind pilot decontamination for massive MIMO systems," *IEEE Transactions on Wireless Communications*, vol. 15, pp. 525–536, Jan 2016.

[6] H. Yin, L. Cottatellucci, D. Gesbert, R. R. Müller, and G. He, "Robust pilot decontamination based on joint angle and power domain discrimination," *IEEE Transactions on Signal Processing*, vol. 64, pp. 2990–3003, June 2016.

[7] J. H. Sørensen and E. de Carvalho, "Pilot decontamination through pilot sequence hopping in massive mimo systems," in *2014 IEEE Global Communications Conference*, pp. 3285–3290, Dec 2014.

[8] H. Yin, D. Gesbert, M. Filippou, and Y. Liu, "A coordinated approach to channel estimation in large-scale multiple-antenna systems," *IEEE J. Sel. Areas Commun.*, vol. 31, no. 2, pp. 264–273, 2013.

[9] A. Adhikary, J. Nam, J. Y. Ahn, and G. Caire, "Joint spatial division and multiplexing - the large-scale array regime," *IEEE Transactions on Information Theory*, vol. 59, pp. 6441–6463, Oct 2013.

[10] M. Li, S. Jin, and X. Gao, "Spatial orthogonality-based pilot reuse for multi-cell massive MIMO transmission," in *2013 International Conference on Wireless Communications and Signal Processing*, pp. 1–6, Oct 2013.

[11] A. Ashikhmin and T. Marzetta, "Pilot contamination precoding in multi-cell large scale antenna systems," in *2012 IEEE International Symposium on Information Theory Proceedings*, pp. 1137–1141, July 2012.

[12] L. Li, A. Ashikhmin, and T. Marzetta, "Pilot contamination precoding for interference reduction in large scale antenna systems," in *2013 51st Annual Allerton Conference on Communication, Control, and Computing (Allerton)*, pp. 226–232, Oct 2013.

[13] E. Björnson, J. Hoydis, and L. Sanguinetti, "Massive MIMO has unlimited capacity," *IEEE Transactions on Wireless Communications*, vol. 17, pp. 574–590, Jan 2018.

[14] A. Osseiran, F. Boccardi, V. Braun, K. Kusume, P. Marsch, M. Maternia, O. Queseth, M. Schellmann, H. Schotten, H. Taoka, H. Tullberg, M. Uusitalo, B. Timus, and M. Fallgren, "Scenarios for 5G mobile and wireless communications: the vision of the metis project," *Communications Magazine, IEEE*, vol. 52, pp. 26–35, May 2014.

[15] E. Dahlman, S. Parkvall, and J. Skold, *4G: LTE/LTE-Advanced for Mobile Broadband*. Academic Press, 1st ed., 2011.

[16] M. S. Ali, E. Hossain, and D. I. Kim, "LTE/LTE-A random access for massive machine-type communications in smart cities," *IEEE Communications Magazine*, vol. 55, pp. 76–83, January 2017.

[17] L. Sanguinetti, A. A. D'Amico, M. Morelli, and M. Debbah, "Random access in uplink massive MIMO systems: How to exploit asynchronicity and excess antennas," in *2016 IEEE Global Communications Conference (GLOBECOM)*, pp. 1–5, Dec 2016.

[18] E. de Carvalho, E. Björnson, J. H. Sørensen, P. Popovski, and E. G. Larsson, "Random access protocols for massive MIMO," *IEEE Communications Magazine*, vol. 55, pp. 216–222, May 2017.

[19] J. H. Sørensen, E. de Carvalho, and P. Popovski, "Massive mimo for crowd scenarios: A solution based on random access," in *2014 IEEE Globecom Workshops (GC Wkshps)*, pp. 352–357, Dec 2014.

[20] G. Liva, "Graph-Based Analysis and Optimization of Contention Resolution Diversity Slotted ALOHA," *Communications, IEEE Transactions on*, vol. 59, pp. 477 –487, february 2011.

[21] C. Stefanovic, P. Popovski, and D. Vukobratovic, "Frameless ALOHA protocol for Wireless Networks," *IEEE Comm. Letters*, vol. 16, pp. 2087–2090, Dec. 2012.

[22] E. de Carvalho, E. Björnson, J. H. Sørensen, E. G. Larsson, and P. Popovski, "Random pilot and data access in massive MIMO for machine-type communications," *IEEE Transactions on Wireless Communications*, vol. 16, pp. 7703–7717, Dec 2017.

[23] E. Bjørnson, E. de Carvalho, J. H. Sørensen, E. G. Larsson, and P. Popovski, "A random access protocol for pilot allocation in crowded massive MIMO systems," *IEEE Transactions on Wireless Communications*, vol. 16, pp. 2220–2234, April 2017.

[24] H. Han, X. Guo, and Y. Li, "A high throughput pilot allocation for M2M communication in crowded massive MIMO systems," *IEEE Transactions on Vehicular Technology*, vol. 66, pp. 9572–9576, Oct 2017.

[25] H. Han, Y. Li, and X. Guo, "A graph-based random access protocol for crowded massive MIMO systems," *IEEE Transactions on Wireless Communications*, vol. 16, pp. 7348–7361, Nov 2017.

[26] L. Liu and W. Yu, "Massive connectivity with massive MIMO - part I: Device activity detection and channel estimation," *https://arxiv.org/abs/1706.06438*, 2017.

[27] F. Rusek, D. Persson, B. K. Lau, E. Larsson, T. Marzetta, O. Edfors, and F. Tufvesson, "Scaling up MIMO: Opportunities and challenges with very large arrays," *Signal Processing Magazine, IEEE*, vol. 30, pp. 40–60, Jan 2013.

[28] E. Paolini, C. Stefanovic, G. Liva, and P. Popovski, "Coded Random Access: How Coding Theory Helps to Build Random Access Protocols," *IEEE Commun. Mag.*, vol. 53, pp. 144–150, June 2015.

[29] M. Luby, "LT Codes," in *Foundations of Computer Science, 2002. Proceedings. The 43rd Annual IEEE Symposium on*, pp. 271–280, 2002.

[30] J. Sørensen, P. Popovski, and J. Østergaard, "Design and analysis of LT codes with decreasing ripple size," *Communications, IEEE Transactions on*, vol. 60, pp. 3191 –3197, november 2012.

[31] M. G. Luby, M. Mitzenmacher, and A. Shokrollahi, "Analysis of Random Processes via And-Or Tree Evaluation," in *Proc. of 9th ACM-SIAM SODA*, (San Francisco, CA, USA), Jan. 1998.

[32] T. Richardson and R. Urbanke, *Modern Coding Theory*. Cambridge University Press, Cambridge, UK, 2007.

[33] C. Stefanovic, M. Momoda, and P. Popovski, "Exploiting Capture Effect in Frameless ALOHA for Massive Wireless Random Access," in *IEEE WCNC 2014*, (Istanbul, Turkey), Apr. 2014.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TWC.2018.2873400, IEEE Transactions on Wireless Communications

Jesper H. Sørensen (M'10) received the B.Sc. in electrical engineering in 2007, the M.Sc. (cum laude) with the maximum grade average in 2009 and the Ph.D. in wireless communications in 2012, all from Aalborg University, Denmark. In second half of 2011 he worked as an intern at Mitsubishi Electric Research Laboratories. From 2012 to 2016 he was a postdoctoral researcher at Aalborg University, where he is currently an assistant professor. His work is in the areas of channel coding, multimedia communication and protocol design.

Elisabeth de Carvalho received a Ph.D. in electrical engineering from Telecom ParisTech, France. After her Ph.D. she was a post-doctoral fellow at Stanford University, USA and then worked in industry in the field of DSL and wireless LAN. Since 2005, she has been is an associate professor at Aalborg University where she has led several research projects in wireless communications. Her main expertise is in signal processing for MIMO communications with recent focus on massive MIMO including channel measurements, channel modeling, beamforming and protocol aspects. She is a coauthor of the text book "A practical guide to the MIMO radio channel".

Čedomir Stefanović (S'04-M'11-SM'17) received his Dipl.-Ing., Mr.-Ing., and Ph.D. degrees in electrical engineering from the University of Novi Sad, Serbia. He is currently an associate professor at the Department of Electronic Systems, Aalborg University, Denmark. In 2014 he was awarded an individual postdoc grant by the Danish Council for Independent Research. He is involved in several national and European Union projects related to the Internet of Things and fifth-generation communications. He is an editor of the IEEE Internet of Things Journal. His research interests include communication theory, wireless and smart grid communications.

Petar Popovski (S'97-A'98-M'04-SM'10-F'16) is a Professor of Wireless Communications with Aalborg University. He received the Dipl. Ing. degree in electrical engineering and the Magister Ing. degree in communication engineering from the "Sts. Cyril and Methodius" University, Skopje, Republic of Macedonia, in 1997 and 2000, respectively, and the Ph.D. degree from Aalborg University, Denmark, in 2004. He has over 300 publications in journals, conference proceedings, and edited books. He holds over 30 patents and patent applications. He received an ERC Consolidator Grant (2015), the Danish Elite Researcher award (2016), IEEE Fred W. Ellersick prize (2016) and IEEE Stephen O. Rice prize (2018). He is currently a Steering Committee Member of IEEE SmartGridComm and previously served as a Steering Committee Member of the IEEE INTERNET OF THINGS JOURNAL. He is currently an Area Editor of the IEEE TRANSACTIONS ON WIRELESS COMMUNICATIONS. His research interests are in the area of wireless communication and networking, and communication/information theory.