Supporting More Active Users for Massive Access via Data-assisted Activity Detection

Xinyu Bian, Yuyi Mao, and Jun Zhang
Department of Electronic and Information Engineering
The Hong Kong Polytechnic University, Hong Kong
Emails: xinyu.bian@connect.polyu.hk, yuyi-eie.mao@polyu.edu.hk, jun-eie.zhang@polyu.edu.hk

***Abstract*—Massive machine-type communication (mMTC) has been regarded as one of the most important use scenarios in the fifth generation (5G) and beyond wireless networks, which demands scalable access for a large number of devices. While grant-free random access has emerged as a promising mechanism for massive access, its potential has not been fully unleashed. Particularly, the two key tasks in massive access systems, namely, user activity detection and data detection, were handled separately in most existing studies, which ignored the common sparsity pattern in the received pilot and data signal. Moreover, error detection and correction in the payload data provide additional mechanisms for performance improvement. In this paper, we propose a data-assisted activity detection framework, which aims at supporting more active users by reducing the activity detection error, consisting of false alarm and missed detection errors. Specifically, after an initial activity detection step based on the pilot symbols, the false alarm users are filtered by applying *energy detection* for the data symbols; once data symbols of some active users have been successfully decoded, their effect in activity detection will be resolved via *successive pilot interference cancellation*, which reduces the missed detection error. Simulation results show that the proposed algorithm effectively increases the activity detection accuracy, and it is able to support $\sim 20\%$ more active users compared to a conventional method in some sample scenarios.**

***Index Terms*—Internet-of-Things (IoT), massive connectivity, grant-free massive access, data-assisted user activity detection, approximate message passing (AMP).**

I. Introduction

The proliferation of the Internet of Things (IoT), such as connected health, smart home, and intelligent manufacturing, is prompting a rapid revolution of wireless communications. In order to support a massive number of connected devices, massive machine-type communications (mMTC) has become one of the three generic services offered by the fifth generation (5G) wireless networks [1]. A unique feature of mMTC is that, while a huge amount of devices are connected, only a proportion of them sporadically become active, normally with a small amount of data to transmit [2].

Nevertheless, uplink access in legacy wireless networks is generally controlled by grant-based access schemes, where each user first transmits a scheduling request to the base station (BS) and cannot start its data transmission until a grant is received. Although the grant-based access schemes reserve dedicated resources for each user that avoids potential collisions, long latency and significant signalling overhead will be incurred with a large number of devices [3], [4].

Grant-free random access, where users can transmit data without waiting for approval from the BS [5], provides a promising solution for mMTC. In its protocols, the BS needs to detect the set of active users and estimate their channel conditions based on the received pilot signal, before performing data reception operations. Due to the vast amount of devices, users can only be assigned with non-orthogonal pilots, which makes it highly challenging for accurate active user identification and channel estimation at the BS. As a result, accommodating the maximum number of active devices with minimum degradation of communication performance is widely acknowledged as one of the most fundamental design considerations for grant-free massive access [3], [6], [7].

Because of the sporadic traffic pattern of the connected devices, detecting the set of active users turns out to be a compressive sensing problem, for which, many efficient algorithms were developed [8]. In [9], a joint user activity detection and data detection algorithm was proposed for grant-free non-orthogonal multiple access (NOMA) by exploiting the temporal correlations of user activities. A similar problem was later revisited using approximate message passing (AMP) and expectation maximization (EM) in [10]. However, these works assume full channel state information (CSI) available at the BSs, which is practicallly infeasible since most of the users are inactive without transmitting their pilots to the BS. Therefore, joint activity detection and channel estimation has attracted significant attentions most recently [11], [12]. In [11], a joint design of activity detection and channel estimation was proposed based on AMP for massive multi-input multi-output (MIMO) systems, and it was shown that the activity detection error can be arbitrarily small in the asymptotic regime. In addition, a user activity detection and channel estimation approach was developed in [12] by leveraging the joint sparsity from both the spatial and frequency domains. This approach obviates the need of knowing the number of devices.

However, prior works on grant-free massive access mostly follow a separate design approach, i.e., the activity pattern and CSI are estimated without incorporating any information encoded in the received data symbols. In this way, it only utilizes the sparse activity pattern from the received pilot signal, which limits the activity detection accuracy and the data transmission reliability. An important but easily neglected

This work was supported by the General Research Fund (Project No. 15207220) from the Hong Kong Research Grants Council.

observation in grant-free random access is that the same user activity pattern replicates in the received data symbols, which can be exploited to improve the activity detection accuracy for accommodating more connected devices. This inspires the design of a data-assisted activity detection framework in this paper, where the false alarm and missed detection error can both be suppressed. It is worthwhile to note that this idea was initially proposed for a single-antenna NOMA-based massive access system [13], which, however, cannot be easily extended for multi-antenna receptions.

In this paper, we endeavor to reduce the activity detection error by leveraging valuable information obtained in data symbols. The proposed data-aided activity detection framework contains three basic modules, namely, an initial estimator, a false alarm corrector and a missed detection corrector. On one hand, to minimize the false alarm error, energy detection is applied in the false alarm corrector to filter inactive users that are incorrectly determined as active. On the other hand, inspired by the successive interference cancellation (SIC) detection, the missed detection corrector progressively increases the sparsity level of the received pilot signal to reduce the probability of missed detection. Simulation results shows that the proposed framework is able to achieve noticeable improvements in terms of both user activity detection accuracy and data detection error. Moreover, about 20% more active users can be supported by the proposed framework in sample scenarios, compared to that achieved by the separate design.

The rest of this paper is organized as follows. We introduce the system model and two basic tasks of grant-free access in Section II. A data-assisted user activity detection framework is developed in Section III. Simulation results are presented in Section IV, and we conclude this paper in Section V.

Notations: We use lower-case letters, bold-face lower-case letters, bold-face upper-case letters, and math calligraphy letters to denote scalars, vectors, matrices, and sets, respectively. Besides, the conjugate transpose of a matrix $\mathbf{M}$ is denoted as $\mathbf{M}^H$ and the complex Gaussian distribution with mean $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$ is denoted by $\mathcal{CN}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. In addition, the indicator function and the Kronecker product are denoted as $\mathbf{1}(\cdot)$ and $\otimes$, respectively. We use $\text{vec}(\cdot)$ to denote the vectorization operator and let $\text{vec}^{-1}(\cdot)$ denote its inverse.

II. System Model

A. Signal Model

We consider an uplink cellular system as shown in Fig. 1, where a large number of mobile users are simultaneously served by a BS. The scenarios where the mobile users have sporadic uplink data traffic (e.g., the IoT and mMTC) are of particular interests, where only a small fraction of the users have data to transmit and become active at each time instant. The active probabilities of different users are assumed to be identical, and they are denoted as p. We denote the set of mobile users as $\mathcal{N} \triangleq \{1, \cdots, N\}$, and use the activity indicator $u_n \in \{0, 1\}$ to represent whether a user is active for transmission, i.e., $u_n = 1$ indicates the user is active and $u_n = 0$ if it is inactive. The set of active users is represented by $\Xi \triangleq \{j \in \mathcal{N} | u_j = 1\}$ with its cardinality denoted as K ($K \leq N$). For simplicity, the BS is assumed to have M receive antennas while each user transmits with a single antenna.

We adopt the quasi-static block fading channel model, where the channel condition remains unchanged within a transmission block spanning T symbol intervals, and changes independently across different coherence blocks. The uplink channel vector from user n to the BS, denoted as $\mathbf{h}_n$, is modeled as $\mathbf{f}_n = \sqrt{\beta_n}\boldsymbol{\alpha}_n, \forall n$, where $\boldsymbol{\alpha}_n$ and β_n stand for the small-scale and large-scale fading coefficients, respectively. Besides, the users are assumed to be static and thus β_n is known at the BS.

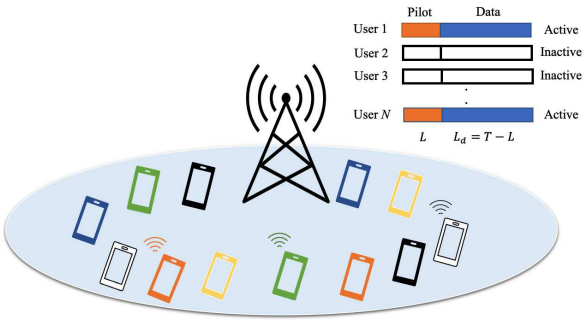


Fig. 1. System model and the adopted grant-free random access scheme.

A grant-free random access scheme as shown in Fig. 1, is adopted for uplink transmissions, where a transmission block is divided into two phases: The first phase contains L symbols that are reserved for pilot transmission and the remaining $L_d \triangleq T - L$ symbols are used for payload data delivery in the second phase. We consider the massive random access scenarios, i.e., $L < N$, in which, assigning orthogonal pilot sequences to all the users is infeasible. To overcome this issue, each user is instead assigned with a unique pilot sequence $\sqrt{L}\mathbf{a}_n$ with $\mathbf{a}_n \triangleq [a_{n,1}, \cdots, a_{n,L}]^T$ and $a_{n,l} \sim \mathcal{CN}\left(0, \frac{1}{L}\right)$ [7]. It can be verified that $\{\mathbf{a}_n\}_{n=1}^N$ achieves asymptotic orthogonality when L is sufficiently large. By defining $\mathbf{A}_p$ as $[\mathbf{a}_1, \cdots, \mathbf{a}_N]$, the received pilot signal $\mathbf{Y}_p \in \mathbb{C}^{L \times M}$ at the BS in the first phase can be expressed as follows:

$$\mathbf{Y}_p = \sqrt{L\rho}\mathbf{A}_p\mathbf{H} + \mathbf{N}_p, \tag{1}$$

where ρ is the user transmit power, $\mathbf{H} \triangleq [\mathbf{h}_1, ..., \mathbf{h}_N]^T$ denotes the effective channel matrix with $\mathbf{h}_n \triangleq u_n\mathbf{f}_n$, and $\mathbf{N}_p = [\mathbf{n}_{p,1}, ..., \mathbf{n}_{p,L}]^T$ is the Gaussian noise with zero mean and variance σ^2 for each element.

In the data transmission phase, each active user transmits s ($s < L_d$) coded symbols, which is denoted as $\mathbf{s}_n \in \mathcal{X}^{s \times 1}$. Here, $\mathcal{X}$ is the set of constellation points with the normalized average power. For the set of inactive users, $\mathbf{s}_n$ is set to be a zero vector for notation consistency. Since the number of active users in the system may far exceed the number of receive antennas at the BS, in order to avoid the system from being overloaded [14], we multiply the coded symbols by a precoding matrix for each user [15] as follows

$$\mathbf{c}_n = \mathbf{P}_n\mathbf{s}_n, \tag{2}$$

where $\mathbf{c}_n$ is the precoded symbols and $\mathbf{P}_n \in \mathbb{C}^{L_d \times s}$ is the precoding matrix with full column-rank. Thus, the received data signal at the BS, denoted as $\mathbf{Y}_d \in \mathbb{C}^{M \times L_d}$, can be expressed as follows:

$$\mathbf{Y}_d = \sqrt{\rho} \sum_{n=1}^{N} \mathbf{h}_n \mathbf{c}_n^T + \mathbf{N}_d = \sqrt{\rho} \sum_{j \in \Xi} \mathbf{h}_j \mathbf{s}_j^T \mathbf{P}_j^T + \mathbf{N}_d, \quad (3)$$

where $\mathbf{N}_d = [\mathbf{n}_{d,1}, ..., \mathbf{n}_{d,L_d}]$ is the Gaussian noise with the same distribution as $\mathbf{N}_p$. We denote $\mathbf{y}_d = \text{vec}(\mathbf{Y}_d)$, and let $\mathbf{B}_n \triangleq \mathbf{P}_n \otimes \mathbf{h}_n$. As a result, the received data signal in (3) can be rewritten as the following expression:

$$\mathbf{y}_d = \sqrt{\rho} \mathbf{B}_a \mathbf{x}_a + \mathbf{N}_d, \quad (4)$$

where $\mathbf{B}_a \triangleq [\{\mathbf{B}_j\}_{j \in \Xi}]$ and $\mathbf{x}_a \triangleq \left[\{\mathbf{s}_j^T\}_{j \in \Xi}\right]$.

B. User Activity and Data Detection

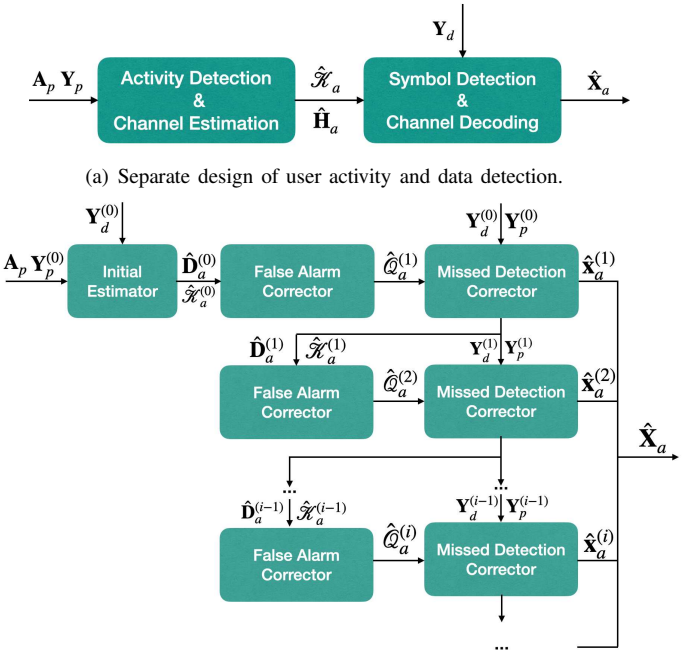
User activity detection and *data detection* are the two most critical tasks in grant-free massive access. Prior studies on massive connectivity typically adopted a two-stage separate design as shown in Fig. 2(a) [10], [16], [17]. Specifically, in the first stage, activity detection and channel estimation are performed based on the received pilot signal, which can be accomplished by exploiting the sparsity of the effective channel matrix using compressive sensing techniques [3]. The estimated user activity pattern and CSI are then used for data detection in the second stage.

With limited resources available for pilot transmissions, it is challenging to obtain accurate knowledge of the user activity pattern at the BS. In fact, missed detection, i.e., an active user is not detected at all, and false alarm, i.e., an inactive user is determined as active, are two major sources that contribute to the user activity detection error. On one hand, data of the miss-detected users is not decoded, leading to a one-hundred percent data error for these users; On the other hand, false alarm shall degrade the data detection accuracy, since the data detector also attempts to decode data for the false alarm users, which is equivalent to introducing interference to the active users. Therefore, improving the activity detection accuracy is of the utmost importance to the communication performance in massive access systems.

A key observation of the grant-free access scheme is that, both the transmitted pilots and data symbols are distorted by the same wireless fading channel. In other words, the received pilot and data signals share the same sparsity pattern, which could be exploited to improve the activity detection accuracy. Nevertheless, this aspect was largely overlooked by existing studies, which motivates our investigation on data-assisted activity detection approaches. In the next section, we will customize dedicated methods to handle the two kinds of errors, in order to reduce the overall user activity detection error for reliable communications.

III. The Proposed Framework

In this section, we propose a data-assisted activity detection framework to improve the activity detection accuracy. A flow chart of the proposed framework is shown in Fig. 2(b), which contains an initial estimator, a false alarm corrector, and a missed detection corrector. For each transmission block, the initial estimator performs preliminary estimation on the CSI and the user activity pattern for subsequent data detection. This is essentially the separate design as shown in Fig. 2(a). Based on the initial data symbol estimates, the false alarm corrector performs energy detection to filter part of the false alarm users. This step is inspired by the intuition that the average magnitudes of the detected data symbols of the false alarm users shall be much smaller than those of the active users. Then, with the updated user activity pattern, the channel matrices and data symbols are re-estimated for processing in the missed detection corrector. In particular, the design of the missed detection corrector leverages the SIC techniques to further refine the activity detection result. In contrast to conventional SIC algorithms that remove interference from the received data signal, interference in the received pilot signal is eliminated by identifying a subset of users whose payload data can be successfully decoded in each iteration.



(a) Separate design of user activity and data detection.

(b) The data-assisted activity detection framework.

Fig. 2. The separate design and data-assisted design for massive access.

We will elaborate different modules of the proposed framework in the following subsections. For better expositions, we use the superscript "$^{(i)}$" to denote the iteration number, and refer the operations of the initial estimator as the 0-th iteration. Besides, the intermediate variables $N^{(0)}$, $K^{(0)}$, $\mathbf{Y}_p^{(0)}$, $\mathbf{y}_d^{(0)}$ and $\mathcal{N}^{(0)}$ are initialized as N, K, $\mathbf{Y}_p$, $\mathbf{y}_d$ and $\mathcal{N}$, respectively.

A. The Initial Estimator

The initial estimator applies the AMP-based algorithm proposed in [10] to jointly estimate the CSI and user activity pattern, based on which, data symbol detection is performed. In

particular, with $\mathbf{Y}_p^{(i)}$ as the input[1], the AMP-based algorithm obtains the channel estimates for the users in $\mathcal{N}^{(i)}$, denoted as $\hat{\mathbf{H}}^{(i)} = \left[\{\hat{\mathbf{h}}_j^{(i)}\}_{j\in\mathcal{N}^{(i)}}\right]$, and the user activity pattern is derived by thresholding, i.e., the set of active users is determined as $\hat{\mathcal{K}}_a^{(i)} \triangleq \left\{j \in \mathcal{N}^{(i)} | \phi(\hat{\mathbf{h}}_j^{(i)}) \geq \theta_j^{(i)}\right\}$, where $\phi(\cdot)$ is a known function and $\theta_j^{(i)}$ is the decision threshold for user j.

We define $\hat{\mathbf{H}}_a^{(i)} \triangleq \left[\{\hat{\mathbf{h}}_j^{(i)}\}_{j\in\hat{\mathcal{K}}_a^{(i)}}\right]$ and $\hat{\mathbf{B}}_a^{(i)}$ in a way similar to $\mathbf{B}_a$ in (4). By using the MMSE equalizer, the estimated data symbols for the users in $\hat{\mathcal{K}}_a^{(i)}$, are obtained via the following expression:

$$\hat{\mathbf{D}}_a^{(i)} = \text{vec}^{-1}\left[\left(\hat{\mathbf{B}}_a^{(i)H}\hat{\mathbf{B}}_a^{(i)} + \frac{\sigma^2}{\rho}\mathbf{I}\right)^{-1}\hat{\mathbf{B}}_a^{(i)H}\mathbf{y}_d^{(i)}\right], \quad (5)$$

where $\hat{\mathbf{D}}_a^{(i)} \triangleq \left[\{\hat{\mathbf{d}}_{a,j}^{(i)}\}_{j\in\hat{\mathcal{K}}_a^{(i)}}\right]$ with $\hat{\mathbf{d}}_{a,j}^{(i)} \triangleq \left[\{\hat{d}_{a,(j,m)}^{(i)}\}_{m=1}^{s}\right]$ and $\hat{d}_{a,(j,m)}^{(i)}$ is the m-th estimated data symbol of user j.

B. The False Alarm Corrector

In the i-th iteration, the false alarm corrector filters the inactive users from $\hat{\mathcal{K}}_a^{(i-1)}$, which is obtained from the initial estimator if $i = 1$ and the missed detection corrector in the previous iteration if $i \geq 2$. We borrow the idea of energy detection for spectrum sensing in cognitive radio networks [18] to design the false alarm corrector. This is because if the estimated data symbols of a user have small average magnitudes, this user is likely to be a false alarm user.

Specifically, in the false alarm corrector, a user that was detected as active in the previous iteration is determined as a false alarm user if the following criteria is satisfied:

$$\sum_{m=1}^{s}\mathbf{1}\left(|\hat{d}_{a,(j,m)}^{(i-1)}| \in (0,\theta_{F_1})\cup(\theta_{F_2},+\infty)\right) \geq \theta_{F_3}, \forall j \in \hat{\mathcal{K}}_a^{(i-1)}. \quad (6)$$

In (6), θ_{F_1}, θ_{F_2} and θ_{F_3} are empirical threshold values, where θ_{F_1} is to ensure the average estimated data symbol energy of an active user is sufficiently large, while θ_{F_2} is designed for reducing the sensitivity of the false alarm corrector to the channel estimation error. Therefore, the updated estimate of the user activity pattern is given by $\hat{\mathcal{Q}}_a^{(i)} \triangleq \hat{\mathcal{K}}_a^{(i-1)} \setminus \{j \in \hat{\mathcal{K}}_a^{(i-1)} | \sum_{m=1}^{s}\mathbf{1}\left(|\hat{d}_{a,(j,m)}^{(i-1)}| \in (0,\theta_{F_1})\cup(\theta_{F_2},+\infty)\right) \geq \theta_{F_3}\}$.

C. The Missed Detection Corrector

*1) **Overview**:* While the false alarm corrector is able to reduce the chances of including inactive users in $\hat{\mathcal{K}}_a^{(i-1)}$, it cannot effectively handle the missed detection users. In this subsection, we design a missed detection corrector to minimize the number of active users that cannot be found in previous steps. Our design is motivated by the SIC techniques for multi-user detection [19], where data from different users are detected sequentially, and interference in the received data signal is iteratively removed for decoding data of the remaining users. However, as our objective is to reduce the missed detection error, we propose to perform SIC for the received pilot signal instead in the missed detection corrector. By identifying some users that are determined as active with high confidence and remove their pilot data from the received pilot signal in each iteration, we shall be able to increase the sparsity level of the received pilot signal, which is beneficial for accurate user activity detection and channel estimation in next iterations.

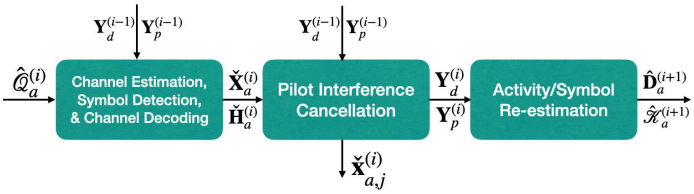


Fig. 3. The structure of the missed detection corrector.

In particular, the missed detection corrector performs three tasks as shown in Fig. 3, including *i) Channel estimation, data symbol detection and channel decoding*; *ii) Pilot interference cancellation*; and *iii) User activity/symbol re-estimation*. These tasks will be elaborated in the sequel.

*2) **Channel estimation, symbol detection, and channel decoding**:* With the updated estimate of the active user set $\hat{\mathcal{Q}}_a^{(i)}$ from the false alarm corrector, the missed detection corrector first re-estimates the channel vectors and performs data symbol detection accordingly, both of which apply the MMSE estimators. The estimated channel vectors of user j is denoted as $\check{\mathbf{h}}_j^{(i)}$, and the detected symbol sequence on the constellation, i.e., which constellation points are transmitted in the symbol sequence, is denoted as $\check{\mathbf{s}}_{a,j}^{(i)}$ for user j. The detected symbol sequence is then passed to a channel decoder, which outputs the parity check result in addition to the decoded data. We denote the set of users that pass the parity check as $\hat{\mathcal{P}}_a^{(i)}$, whose channel-decoded data is denoted as $\check{\mathbf{x}}_{a,j}^{(i)}$, $j \in \hat{\mathcal{P}}_a^{(i)}$.

*3) **Pilot interference cancellation and activity/symbol re-estimation**:* After channel decoding, the missed detection corrector selects a number of users from $\hat{\mathcal{Q}}_a^{(i)}$ based on their parity check results. The pilots of the selected set of users are then subtracted from the received pilot signal. Our heuristics originate from a key theorem in compressive sensing (See Theorem 1.3 in [20]). This theorem implies that for an idealized user activity detection problem where the BS has a single receive antenna and without the receive noise, if t ($t \leq K$) of the active users can be identified by an oracle, the remaining $K-t$ active users can also be accurately identified from the interference-cancelled pilot signal as long as the triplet (N, K, L) satisfies the following inequality:

$$L \geq C(K-t)\ln\left(\frac{N-t}{K-t}\right), t = 0, 1, 2, \cdots, K, \quad (7)$$

where $C > 0$ is a constant. Since the right-hand side of (7) decreases with t, it means that if more active users can be

[1]In this subsection, we retain the superscript "$^{(i)}$" as the key steps in the initial estimator that are reused in the missed detection corrector as will be discussed in Section III-D, in which the iteration number is greater than zero.

accurately found by an oracle, the perfect active user recovery condition can be met more easily for the remaining users. In other words, increasing the sparsity level in the received pilot signal is useful to improve the activity detection performance. Unfortunately, this conclusion is drawn by imposing strict assumptions, which is rarely the case in practice.

To resolve this issue, the missed detection corrector selects $\min\{S_a, |\hat{\mathcal{P}}_a^{(i)}|\}$ users from $\hat{\mathcal{Q}}_a^{(i)}$ based on the channel decoding results, and the set of selected users for pilot interference cancellation in the i-th iteration, is denoted as $\hat{\mathcal{M}}_a^{(i)}$. Here, S_a is a preset parameter in the proposed framework. In case that $|\hat{\mathcal{P}}_a^{(i)}| > S_a$, the S_a users with the minimum Euclidean distances between $\check{\mathbf{x}}_{a,j}^{(i)}$ and $\check{\mathbf{s}}_{a,j}^{(i)}$ are selected. Thereafter, the pilot data of the selected set of users is subtracted from the received pilot signal $\mathbf{Y}_p^{(i-1)}$ for the next iteration:

$$\mathbf{Y}_p^{(i)} = \mathbf{Y}_p^{(i-1)} - \sqrt{L\rho} \sum\nolimits_{j \in \hat{\mathcal{M}}_a^{(i)}} \check{\mathbf{h}}_j^{(i)} \mathbf{a}_j^T, \tag{8}$$

and $\mathbf{Y}_d^{(i)}$ is updated accordingly as follows:

$$\mathbf{Y}_d^{(i)} = \mathbf{Y}_d^{(i-1)} - \sqrt{\rho} \sum\nolimits_{j \in \hat{\mathcal{M}}_a^{(i)}} \check{\mathbf{x}}_{a,j}^{(i)} \check{\mathbf{h}}_j^{(i)}. \tag{9}$$

If $\hat{\mathcal{P}}_a^{(i)} = \emptyset$, the proposed algorithm will be terminated and the data decoding results of all users from $\hat{\mathcal{Q}}_a^{(i)}$ are deemed as incorrect. Otherwise, the AMP-based algorithm adopted by the initial estimator will be invoked again with $\mathbf{Y}_p^{(i)}$, $\mathbf{Y}_d^{(i)}$, $N^{(i)} = N^{(i-1)} - |\hat{\mathcal{M}}_a^{(i)}|$ and $K^{(i)} = K^{(i-1)} - |\hat{\mathcal{M}}_a^{(i)}|$ as input for re-estimating the user activity pattern and data symbols before calling the false alarm corrector in the next iteration.

IV. Simulation Results

A. Simulation Settings and Baseline Schemes

We simulate a single-cell uplink cellular network with $N = 500$ users to corroborate the effectiveness of the proposed data-assisted user activity detection algorithm, where the users are located on a circle with a radius of 500 m to the BS. Each element of the precoding matrix $\mathbf{P}_n$ is sampled from the complex Gaussian distribution with zero mean and unit variance. In addition, we apply an idealized channel coding scheme, where perfect data recovery is assumed to be feasible if the symbol error in each block is below 20%. It is worthy mentioning that the proposed algorithm is readily applicable for practical channel coding schemes, such as the low-density parity-check codes (LDPC) [21]. The simulation results are averaged over 10^7 independent realizations. Other critical parameters used in simulations are summarized in TABLE I.

TABLE I
Simulation parameters

Parameters	Values	Parameters	Values
M	32	L	100
L_d	70	s	5
β_n	-116.78 dB	$\boldsymbol{\alpha}_n$	$\mathcal{CN}(\mathbf{0}, \mathbf{I}_M)$
θ_{F_1}	0.2	θ_{F_2}	2
θ_{F_3}	3	S_a	20
Bandwidth	1 MHz	Modulation	QPSK
Transmit power	23 dBm	Noise power density	-169 dBm/Hz

We adopt three baseline schemes for comparisons:

- **Separate design:** This scheme was proposed in [10], where channel estimation and user activity detection are first performed using the AMP-based algorithm as stated in Section III-A. After that, data symbols are detected using an MMSE estimator.
- **The proposed algorithm with false alarm correction only:** The main purpose of comparing this scheme is to reveal the impacts of the false alarm users on the system performance, including the activity detection error and data error. Specifically, we execute the proposed framework without invoking the missed detection corrector in this baseline scheme.
- **Perfect knowledge of the user activity pattern:** This scheme assumes perfect knowledge of the user activity pattern and consequently, channel estimation and data detection can be performed for the active users as those in conventional uplink cellular networks. However, as the user activity pattern cannot be known as prior, this scheme is unachievable in practice but can serve as a valuable performance upper bound.

B. Results

We first evaluate the user activity detection error rate, including the false alarm and missed detection probabilities, depicted in Fig. 4. As seen from this figure, both the false alarm and missed detection probabilities increase with the number of active users. This is owing to the limited pilot resources available for user activity detection. Besides, for both the proposed framework and the separate design, it is observed that the false alarm probabilities dominate, which confirms the significance of the false alarm users to the overall activity detection accuracy. In addition, compared to the separate design, the proposed framework drastically reduces both types of activity detection error, which verifies its effectiveness in improving the activity detection performance by fully utilizing the sparsity pattern encoded in the received pilot and data signals. Moreover, we see that the performance improvement achieved by the proposed framework compared to the separate design is much more remarkable when K is below 100, indicating that it is most effective when the traffic load in the systems ranges from light to medium.

Next, we investigate the data error performance achieved by different algorithms and show the relationship between the block error rates (BLERs) and the number of active users in Fig. 5. Similar to the user activity detection error, the BLERs increase with the number of active users in the system. It is also noticed that the false alarm users have a significant impact on the BLER performance. For instance, when $K = 100$, the proposed framework is able to reduce the BLER from 7×10^{-3} to 4×10^{-3} by invoking the false alarm correction only, while further applying the missed detection corrector only secures an extra 14.3% BLER reduction. This matches

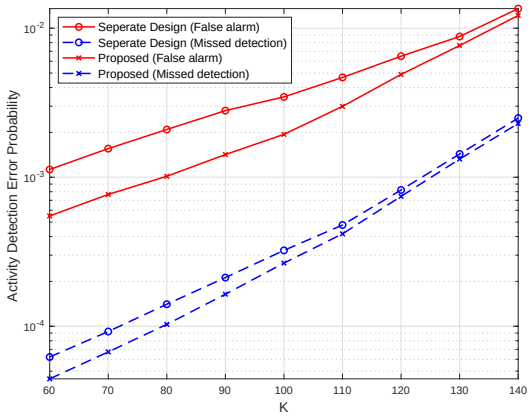


Fig. 4. Activity detection error probability vs. the number of active users.

the results in Fig. 4, where false alarm dominates the activity detection error. In addition, our proposed framework is able to support a substantially larger amount of active users compared to the baselines. For instance, if the BLER requirement is set to be 10^{-3}, the proposed data-assisted user activity detection algorithm is capable of supporting fifteen additional users, which is a more than 20%-improvement compared to the separate design. This again validates the superiority of the data-assisted design by fully exploiting the signal sparsity.

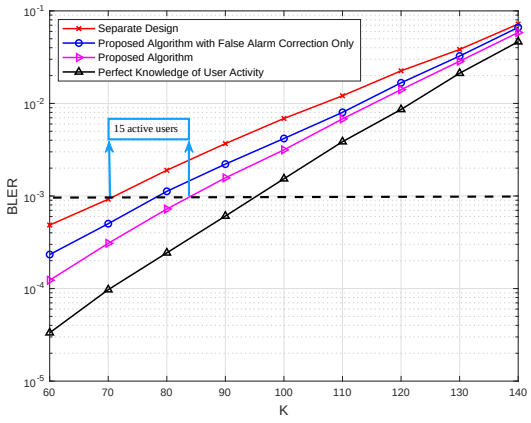


Fig. 5. BLER vs. the number of active users.

V. Conclusions

In this paper, we proposed a data-assisted user activity detection framework for massive random access. This framework effectively exploits the common sparsity pattern in both the received pilot and data signal, and thus boosts the performance of massive access for mMTC applications. Simulation results demonstrated that with the proposed framework, more than 20% of active users can access the network with sufficient reliability. Based on this promising result, we advocate for a holistic approach on designing massive random access systems, by integrating the tasks of activity detection, channel estimation, and data detection and fully exploiting the available prior structure information. This calls for further investigations on efficient algorithms and theoretical analysis.

References

[1] Y. Shi, J. Dong, and J. Zhang, *Low-overhead Communications in IoT Networks - Structured Signal Processing Approaches*, Springer, 2020.

[2] C. Bockelmann *et al.*, "Massive machine-type communications in 5G: Physical and MAC-layer solutions," *IEEE Commun. Mag.*, vol. 54, no. 9, pp. 59–65, Sep. 2016.

[3] L. Liu *et al.*, "Sparse signal processing for grant-free massive connectivity: A future paradigm for random access protocols in the Internet of Things," *IEEE Signal Process. Mag.*, vol. 35, no. 5, pp. 88-99, Sep. 2018.

[4] P. Schulz *et al.*, "Latency critical IoT applications in 5G: Perspective on the design of radio interface and network architecture," *IEEE Commun. Mag.*, vol. 55, no. 2, pp. 70-78, Feb. 2017.

[5] M. T. Islam, A. E. M. Taha, and S. Akl, "A survey of access management techniques in machine type communications," *IEEE Commun. Mag.*, vol. 52, no. 4, pp. 74–81, Apr., 2014.

[6] S. Jiang, X. Yuan, X. Wang, C. Xu and W. Yu, "Joint user identification, channel estimation, and signal detection for grant-free NOMA," *IEEE Trans. Wireless Commun.*, vol. 19, no. 10, pp. 6960-6976, Oct. 2020.

[7] Y. Wu *et al.*, "Massive access for future wireless communication systems," *IEEE Wireless Commun.*, vol. 27, no. 4, pp. 148-156, Aug. 2020.

[8] D. L. Donoho, "Compressed sensing," *IEEE Trans. Inf. Theory*, vol. 52, no. 4, pp. 1289–1306, Apr. 2006.

[9] B. Wang, L. Dai, Y. Zhang, T. Mir, and J. Li, "Dynamic compressive sensing-based multi-user detection for uplink grant-free NOMA," *IEEE Commun. Lett.*, vol. 20, no. 11, pp. 2320–2323, Nov. 2016.

[10] C. Wei, H. Liu, Z. Zhang, J. Dang, and L. Wu, "Approximate message passing-based joint user activity and data detection for NOMA," *IEEE Commun. Lett.*, vol. 21, no. 3, pp. 640–643, Mar. 2017.

[11] L. Liu and W. Yu, "Massive connectivity with massive MIMO–Part I: Device activity detection and channel estimation," *IEEE Trans. Signal Process.*, vol. 66, no. 11, pp. 2933–2946, Jun. 2018.

[12] M. Ke, Z. Gao, Y. Wu, X. Gao, and R. Schober, "Compressive sensing based adaptive active user detection and channel estimation: Massive access meets massive MIMO," *IEEE Trans. Signal Process.*, vol. 68, pp. 764–779, 2020.

[13] Y. Du *et al.*, "Joint channel estimation and multiuser detection for uplink grant-free NOMA," *IEEE Wireless Commun. Lett.*, vol. 7, no. 4, pp. 682–685, Feb. 2018.

[14] R. Miguel, V. Gardašević, R. Müller, and F. Knudsen, "On overloaded vector precoding for single-user MIMO channels," *IEEE Trans. Wireless Commun.*, vol. 9, no. 2, pp. 745–753, Feb. 2010.

[15] R. Xie, H. Yin, Z. Wang, X. Chen and G. Wei, "Many access for small packets based on precoding and sparsity-aware recovery," in *Proc. IEEE Global Commun. Conf (GLOBECOM)*, Washington, DC, USA, Dec. 2016.

[16] G. Wunder, P. Jung, and C. Wang, "Compressive random access for post-LTE systems," in *Proc. IEEE Int. Conf. Commun. (ICC) Workshops*, Sydney, Australia, Jun. 2014.

[17] H. F. Schepker, C. Bockelmann, and A. Dekorsy, "Exploiting sparsity in channel and data estimation for sporadic multi-user communication," in *Proc. Int. Symp. Wireless Commun. Syst. (ISWCS)*, Ilmenau, Germany, Aug. 2013.

[18] W. Zhang, R. K. Mallik and K. B. Letaief, "Optimization of cooperative spectrum sensing with energy detection in cognitive radio networks," *IEEE Trans. Wireless Commun.*, vol. 8, no. 12, pp. 5761-5766, Dec. 2009.

[19] D. Tse and P. Viswanath, *Fundamentals of Wireless Communication*. Cambridge University Press, 2005.

[20] S. Foucart and H. Rauhut, *A Mathematical Introduction to Compressive Sensing*. Springer, 2013.

[21] R. G. Gallager, "Low density parity check codes," *IRE Trans. Inf. Theory*, vol. IT-8, no. 1, pp. 21-28, Jan. 1962.