

A Joint Activity and Data Detection Scheme for Asynchronous Grant-Free Rateless Multiple Access

Wei Zhang*, Xiaofeng Zhong, Shidong Zhou

Department of Electronic Engineering, Tsinghua University, Beijing 100084, China

* The corresponding author, email: zhang-w18@mails.tsinghua.edu.cn

Cite as: W. Zhang, X. Zhong, *et al.*, "A joint activity and data detection scheme for asynchronous grant-free rateless multiple access," *China Communications*, 2024, vol. 21, no. 1, pp. 34-52. DOI: 10.23919/JCC.fa.2023-0355.202401

Abstract: This paper considers the frame-asynchronous grant-free rateless multiple access (FA-GF-RMA) scenario, where users can initiate access at any symbol time, using shared channel resources to transmit data to the base station. Rateless coding is introduced to enhance the reliability of the system. Previous literature has shown that FA-GF-RMA can achieve lower access delay than frame-synchronous grant-free rateless multiple access (FS-GF-RMA), with extreme reliability enabled by rateless coding. To support FA-GF-RMA in more practical scenarios, a joint activity and data detection (JADD) scheme is proposed. Exploiting the feature of sporadic traffic, approximate message passing (AMP) is exploited for transmission signal matrix estimation. Then, to determine the packet start points, a maximum posterior probability (MAP) estimation problem is solved based on the recovered transmitted signals, leveraging the intrinsic power pattern in the codeword. An iterative power-pattern-aided AMP algorithm is devised to enhance the estimation performance of AMP. Simulation results verify that the proposed solution achieves a delay performance that is comparable to the performance limit of FA-GF-RMA.

Keywords: asynchronous grant-free; JADD; rateless codes

I. INTRODUCTION

1.1 Backgrounds and Motivations

The proliferation of connected devices in the Internet of Things (IoT) [1] has given rise to numerous emerging applications of massive Machine Type Communications (mMTC) [2]. In the era of the Internet of Things, wireless networks will provide differentiated services [3], including wide coverage, large-scale connectivity, high reliability, low latency, and high data rates for a vast number of devices. In the context of mMTC, data traffic exhibits characteristics such as a high proportion of uplink data, and bursty short packets [4]. These features drive the need for an efficient random access scheme to address various service requirements, including large-scale access, low power usage, and Ultra-Reliable and Low-Latency Communications (URLLC) [5].

Conventional grant-based access procedure, adopted in Long Term Evolution-Advanced (LTE-A) [6] and the fifth-Generation New Radio (5G NR) [7], requires a four-step handshaking protocol, through which orthogonal radio resources are allocated to different users for uplink transmission. However, such protocol is inefficient in mMTC scenario, since the payload size may be much smaller than the signaling overhead. Moreover, it is hard to support massive device connectivity due to the frequent preamble collision [8].

To support URLLC and mMTC, grant-free access

Received: Jun. 12, 2023
Revised: Oct. 25, 2023
Editor: Yifei Yuan

has been proposed and extensively discussed in both industry and academia. In this approach, the traditional grant request before uplink transmission is omitted to reduce control signaling overhead and access latency [9].

However, due to the lack of scheduling coordination, grant-free access may encounter serious collision problems where multiple users simultaneously use the same channel resources. To solve this problem, advanced receivers in Non-Orthogonal Multiple Access (NOMA) can recover multiple user data streams from interfered signals [10, 11]. In the context of NOMA, grant-free access faces a variable number of interfering users, leading to more dynamic channel conditions [12]. To enhance reliability, conventional link adaptation methods select a fixed-rate code based on channel state information (CSI), involving feedback that results in additional waiting delays [13]. In contrast, rateless coding offers an efficient, flexible alternative, which dynamically sets the code rate during transmission, adapting to CSI changes on the fly [14]. Rateless coding involves the continuous generation of encoded symbols by the sender, without waiting for feedback. The receiver decodes the message successfully once it collects a sufficiently long codeword [15]. Common rateless codes for additive white Gaussian noise (AWGN) channels include Analog Fountain Codes (AFC) [16] and Spinal codes [17]. Notably, recent research on short packet AFC [18–20] has unveiled AFC's potential to achieve latencies comparable to the Polyanskiy-Poor and Verdú (PPV) normal approximation bound [21], indicating AFC as a promising solution for URLLC. In this paper, we term grant-free access with rateless codes as grant-free rateless multiple access (GF-RMA).

Existing research on grant-free access typically assumes frame-synchronous transmission, restricting users to start access only at the beginning of a transmission period or frame, leading to potential waiting delays. In frame-synchronous grant-free rateless multiple access (FS-GF-RMA), without strict delay constraints but zero tolerance for packet loss, the transmission period continues until all active users are successfully decoded. If there is a high number of active users or a user experiences poor channel conditions, the transmission period could extend significantly. Consequently, a substantial group of users with new packets may initiate their access in unison in the following

transmission period, potentially further lengthening it. As demonstrated in [22], an unstable condition arises in the FS-GF-RMA system when the packet arrival rate is excessively high. In such cases, the average number of active users during the transmission period surpasses that of the previous period, and the average transmission delay increases over time to infinity.

To address this, frame-asynchronous grant-free rateless multiple access (FA-GF-RMA) is proposed in [23], which allows users to initiate access at any time, thus eliminating waiting delays. The authors in [24] show that FA-GF-RMA is stable and achieves significantly lower delays than FS-GF-RMA, especially in low signal-to-noise ratio (SNR) or high packet arrival rate scenarios, indicating that FA-GF-RMA is a promising approach for achieving URLLC in power-constrained or high-traffic-load conditions.

However, the delay comparison in [24] is conducted under ideal conditions including perfect knowledge of activity and channel state, as well as capacity-achieving codes. Whether the delay advantages of FA-GF-RMA persist in more practical conditions needs to be verified. Furthermore, to fulfill the advantages of FA-GF-RMA, it is necessary to address the potential challenges in its practical implementation, including activity detection, transmission starting point estimation, channel estimation, and demodulation.

1.2 Related Works

In scenarios with low mobility, such as indoor environments or monitoring with fixed sensors, the channel coherence time greatly exceeds the data transmission time. In such cases, it is feasible to assume that the receiver has prior knowledge of CSI for each potential device, which can be obtained through an initial channel estimation process. The receiver's task is to perform joint activity and data detection (JADD) over an extended coherent period. In this work, we consider the case of perfect CSI at the receiver.

Due to the sporadic feature of traffic in mMTC, JADD is often cast as a sparse signal recovery problem, which can be efficiently solved by compressive sensing (CS) techniques. A great number of works have studied CS-based JADD for synchronous grant-free access.

In [25], the temporal correlation of the active user sets between adjacent time slots is exploited to realize

JADD in several continuous time slots through the Orthogonal Matching Pursuit (OMP) algorithm. While in [26], a parameter evaluating the quality of the previous active user set is introduced, which is used as prior information to aid the subspace pursuit (SP) algorithm for solving JADD. The greedy-based algorithms fail to utilize the prior distribution of the transmitted signals, and the matrix inversions incorporated in them cause high computational complexity when the system scale is large. Recently, the CS techniques under the framework of Bayesian inference, such as Approximate Message Passing (AMP) [27], Orthogonal Approximate Message Passing (OAMP) [28], Expectation propagation (EP) [29] and so on, have gained great attention in JADD due to their flexibility and accuracy. In [30], an AMP-based detection scheme is proposed with the hyper-parameters of device activity learned through the Expectation Maximization (EM) algorithm. The authors in [31] develop an OAMP-based algorithm to solve the multiple measurement vectors (MMV) problem, which devises a structure-learning module to leverage the structured sparsity of uplink signals among multiple time slots.

Some research works have explored frame-asynchronous access without uplink synchronization. Assuming that timing offsets are integer multiples of the symbol interval, [32] proposes a model-driven deep learning approach, which is termed Learned Approximate Message Passing Algorithm (LAMP) to address issues concerning activity detection, time offset estimation, and channel estimation. The authors in [33] consider scenarios with continuous-valued timing offsets where inter-symbol-interference exists. By assuming the knowledge of the delay profile at the receiver, they devise a sparse Bayesian learning algorithm based on the message-passing framework for activity detection and channel estimation, which allows the receiver to harness timing offsets among users for improved multi-user interference suppression.

It's worth noting that in frame-synchronous access without uplink synchronization, the starting points of user data packets are also misaligned but within the maximum propagation delay of the cell. While in the frame-asynchronous grant-free access considered in this paper, users can initiate access at any time, and the deviations between the starting points of user data packets may be comparable to the packet length.

Some researchers have investigated the blind detection problem of frame-asynchronous grant-free access. In [34], the sliding-window detection framework is employed, and by exploiting the sparsity structure embedded in the signal matrix in an observation window, a blind detection algorithm termed turbo bilinear generalized approximate message passing (Turbo-BiG-AMP) is proposed to recover the channel, the transmission start points of active users, and the user data in a single phase. In [35], the Turbo-BiG-AMP algorithm is further extended to asynchronous grant-free access without uplink synchronization. However, both studies assume the receiver knows the number of active users in an observation window, which might limit their applicability in grant-free access scenarios where the number is often unknown in advance.

The above-mentioned works all address receiver design issues for grant-free access with a fixed data rate. As for frame-asynchronous grant-free access with rateless codes, the authors in [36] have devised a novel joint user activity detection and channel estimation scheme. However, the delay performance of FA-GF-RMA with practical rateless codes is not demonstrated in [36]. In this paper, we focus on the joint detection of user activities and data with the knowledge of CSI at the receiver and evaluate the delay performances of FA-GF-RMA employing AFC.

1.3 Our Contribution

This paper investigates the JADD problem for FA-GF-RMA. To the best of our knowledge, this is the first attempt to address the JADD problem in the FA-GF-RMA system. To this end, we utilized the AMP-based JADD algorithm to recover the active user set and their transmitted signals, and then estimate the transmission start points by exploiting the intrinsic power mode. The proposed JADD solution can be used in FA-GF-RMA with any specific rateless code. Particularly, we use AFC as an example to verify the delay performances of the proposed JADD solution.

The rest of the paper is organized as follows. Section II describes the system model of FA-GF-RMA and briefly introduces the AFC. Section III formulates the JADD problem in FA-GF-RMA. In Section IV, we present the proposed solution. Simulation results are given in Section V to demonstrate the effectiveness of the proposed algorithm. Finally, conclusions are

drawn in Section VI.

Notations: We use bold upper and lower case letters for matrices and column vectors, respectively. For a matrix $\mathbf{S}$, we use $s_{i,j}$, $\mathbf{s}_j$, $\mathbf{S}_{k,:}$ and $\mathbf{S}_{k,a:b}$ to denote the entry in the i -th row and the j -th column, the j -th column, the k -th row and the row vector composed of its a -th to b -th elements in the k -th row, respectively. $\mathbf{I}_M$ denotes the identity matrix of size $M \times M$. $\mathcal{CN}(\mathbf{x}; \boldsymbol{\mu}, \boldsymbol{\Gamma})$ denotes the complex Gaussian distribution of a random vector $\mathbf{x}$ with mean $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Gamma}$. $\mathbb{R}^+$ and $\mathbb{C}$ represent the positive real number and complex number set, respectively.

II. SYSTEM MODEL

We consider a typical scenario of uplink grant-free access. A base station equipped with G antennas serves K Internet of Things (IoT) devices. Assume that all users have achieved uplink synchronization with the base station. Each user generates a fixed-size data packet stream according to a Poisson process with an arrival rate of λ packets per symbol. In FA-GF-RMA, users are allowed to initiate data packet transmission at the beginning of any symbol interval.

When a user's data packet arrives, it is first encoded via AFC to generate a potentially infinite-length codeword. Then the coded symbols are transmitted successively until an acknowledgment (ACK) is received from the base station or the maximum transmission time is reached¹. If an ACK is not received before reaching the maximum transmission time L , the packet is considered an outage.

Let $\tau_{k,j}$ denote the time at which user k starts transmitting its j -th data packet, and $\zeta_{k,j}$ denote the time at which user k stops transmitting its j -th data packet. As shown in Figure 1, the first packet of user 1 starts transmission at $\tau_{1,1}$ and receives an ACK at $\zeta_{1,1}$, resulting in an early termination of the transmission. While the first packet of user 2 starts transmission at $\tau_{2,1}$ but continues until the maximum codeword length is reached without receiving an ACK, therefore, it is outaged.

2.1 Analog Fountain Codes

The analog fountain code (AFC) is a representative rateless code, initially introduced in [16]. In this work, we choose AFC as a specific rateless coding scheme to exemplify the demodulation and decoding challenges

of FA-GF-RMA, considering the AFC's significant promise in achieving URLLC. For the convenience of our readers, we first provide a brief overview of AFC's encoding and decoding processes.

Before applying AFC encoding, the entire data block of N_b information bits undergo binary phase shift keying (BPSK) modulation to obtain N_b information symbols $\mathbf{b} \triangleq [b_1, \dots, b_{N_b}]^T \in \{1, -1\}^{N_b \times 1}$.

The encoding process of AFC is represented by an infinite-row generator matrix $\mathbf{G}$, featuring a constant number of columns denoted as N_b . Each row of matrix $\mathbf{G}$ is inherently sparse, with its non-zero elements uniformly drawn from a pre-defined set $\mathcal{W}$ of real-number weights, evenly distributed within the row. The number of non-zero elements, denoted as d_i for the i -th row, is determined by a pre-defined degree distribution function $\Omega(x) = \sum_{i=1}^k \Omega_i x^i$. Denote the index of the j -th non-zero element in the i -th row as $I_{i,j}^b$, and the corresponding index of weight coefficient as $I_{i,j}^w$, then the i -th real-domain AFC coded symbol $\bar{c}_i$ is generated as

$$\bar{c}_i = \sum_{j=1}^{d_i} w_{I_{i,j}^w} b_{I_{i,j}^b}. \quad (1)$$

Regarding the coded symbols as factor nodes and information symbols as variable nodes, the AFC encoding process can also be described by a bipartite graph, as shown in Figure 2. Assuming the weight set is energy-normalized, i.e., $\sum_{i=1}^{|\mathcal{W}|} w_i^2 = 1$, the average power of the i -th AFC-coded symbol d_i .

To fully utilize the constellation plane, each two consecutive real-domain coded symbols compose one complex-domain symbol. Specifically, define a modulator Φ , that transforms the real-domain coded symbols into complex-domain symbols:

$$\Phi([\bar{c}_1, \dots, \bar{c}_{2n}]^T) = [c_1, \dots, c_n]^T, \quad (2)$$

where $c_i \triangleq \bar{c}_{2i-1} + \sqrt{-1}\bar{c}_{2i}$.

Theoretically, the AFC encoder can generate an infinite number of coded symbols. Practically, considering the maximum transmission time L , each packet may be encoded into at most $2L$ real-domain coded symbols. The generator matrix $\mathbf{G}$ is of size $2L \times N_b$.

Upon receiving n_0 ($n_0 < 2L$) real-domain AFC coded symbols, the receiver attempts to decode. If the decoding fails, the receiver continues to collect additional δ symbols and reruns the decoder. This

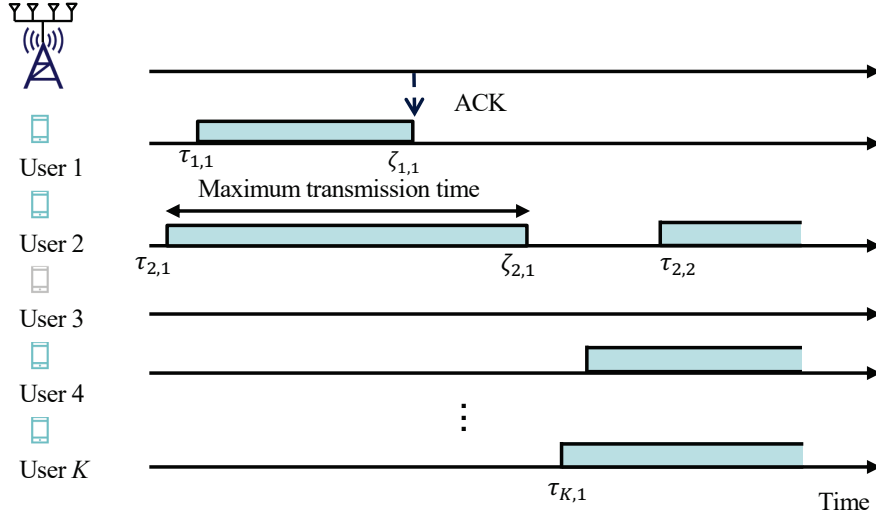


Figure 1. The sporadic and asynchronous packet transmissions in FA-GF-RMA.

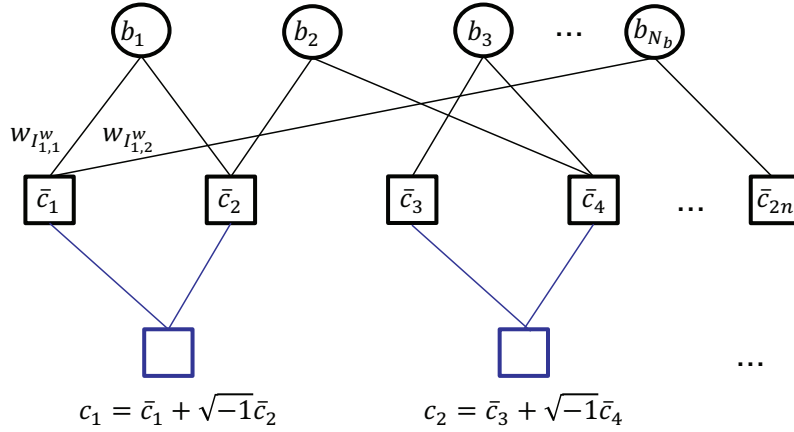


Figure 2. The encoding bipartite graph of AFC and the modulation of real-domain coded symbols.

process continues until successful decoding or the received codeword reaches the maximum length. At the stage of receiving n real-domain coded symbols, the receiver constructs a sub-graph based on the 1st to the n -th row of $\mathbf{G}$ and then runs the belief propagation (BP) decoding algorithm on the graph to obtain the log-likelihood ratios (LLRs) for each information symbol [37]. To mitigate error floor, in practice, a fixed-rate precoder may be concatenated to the AFC encoder, e.g. a low-density parity check precoder.

In a multi-user access scenario, each user k is allocated a unique random seed to construct the genera-

tor matrix $\mathbf{G}_k$. The assignment of these random seeds is known to the receiver, enabling the retrieval of the generator matrices on the receiver's side for decoding.

2.2 Signal Model

Let $\mathbf{h}_k$ denote the channel vector of user k , following a complex Gaussian distribution $\mathcal{CN}(0, \beta_k \mathbf{I}_G)$, where β_k represents the path loss and large-scale fading and G is the number of antennas.

Let $\mathbf{c}_{k,j} = [c_{k,j}(1), \dots, c_{k,j}(L)]^T \in \mathbb{C}^{L \times 1}$ represent the AFC codeword of the j -th data packet of user k , and $s_{k,t}$ represents the signal transmitted by user k at

time t . Then

$$s_{k,t} = \begin{cases} \sqrt{p_{k,l}} c_{k,j}(l), & \exists j, \tau_{k,j} \leq t \leq \zeta_{k,j}, \\ 0, & \text{otherwise,} \end{cases} \quad (3)$$

where $l = t - \tau_{k,j} + 1$ and $p_{k,l}$ is the power gain factor of the l -th coded symbol of user k 's packets. The average transmission power $v_{k,l}^2$ of the l -th AFC-coded symbol of user k is given by

$$v_{k,l}^2 = p_{k,l}(d_{k,2l-1} + d_{k,2l}), \quad (4)$$

where $d_{k,l}$ is the degree of the l -th real-domain AFC-coded symbol of user k .

At the t -th symbol instant, the received signal at the base station can be expressed as

$$\mathbf{y}_t = \mathbf{H}\mathbf{s}_t + \mathbf{n}_t \in \mathbb{C}^{G \times 1}, \quad (5)$$

where $\mathbf{H} = [\mathbf{h}_1, \dots, \mathbf{h}_K]$ represents the matrix composed of channel vectors of for K potential users, G is the number of receive antennas, $\mathbf{s}_t = [s_{1,t}, \dots, s_{K,t}]^T$ represents the vector of transmitted signals from K users at time t , and $\mathbf{n}_t$ represents the Gaussian white noise distributed as $\mathcal{CN}(\mathbf{0}, \sigma^2 \mathbf{I}_G)$.

This paper assumes that the channel matrix $\mathbf{H}$ is known at the receiver. To accomplish decoding, the receiver needs to recover the starting points of each data packet $\{\tau_{k,j} | \forall k, j\}$ and the data packet codewords $\{c_{k,j} | \forall k, j\}$ based on the received signals $\{\mathbf{y}_t | t \geq 0\}$.

2.3 The Sliding-Window Framework

Given user-initiated access at arbitrary times, we employ a sliding-window framework for packet detection, with window length denoted as W and the sliding step as Δ . While a larger W enhances detection performance through extended observations, it concurrently increases complexity. Given that the maximum codeword length of each packet is L , it is judicious to set $W = L$. This configuration allows the receiver to observe the maximum-length codeword at a minimal increase in computational complexity.

Assuming the current time is t , then the sliding window spans from $t - L + 1$ to t . Within this window,

the signal model is denoted as

$$\mathbf{Y}^{(t)} = \mathbf{H}\mathbf{S}^{(t)} + \mathbf{N}^{(t)}, \quad (6)$$

where $\mathbf{S}^{(t)} = [\mathbf{s}_{t-L+1}, \dots, \mathbf{s}_t] \in \mathbb{C}^{K \times L}$, and $\mathbf{N}^{(t)} = [\mathbf{n}_{t-L+1}, \dots, \mathbf{n}_t] \in \mathbb{C}^{G \times L}$.

Active users, in this context, are defined as those who transmit signals within the window. As depicted in Figure 3, there are three distinct types of packets transmitted by active users within the sliding window:

- Type I: packets that start transmission before $t - L + 1$ and stop within the window, possibly due to successful decoding or an outage event.
- Type II: packets that initiate and terminate transmission during the window, signifying successful decoding.
- Type III: packets that start transmission within the window but remain undecoded at this stage.

In each window, it is only necessary to detect type III packets. Let $\mathcal{A}_t$, $\mathcal{D}_t$, and $\mathcal{S}_t$ represent, respectively, the sets of active users, the starting points, and the transmitted signals corresponding to type III packets in the t -th window, which will be formulated in the next section.

As illustrated in Figure 3, the JADD module provides the estimation of $\mathcal{A}_t$, $\mathcal{D}_t$ and $\mathcal{S}_t$, denoted as $\hat{\mathcal{A}}_t$, $\hat{\mathcal{D}}_t$ and $\hat{\mathcal{S}}_t$. Subsequently, the decoder bank concurrently attempts to decode the packets of users in $\hat{\mathcal{A}}_t$ based on the estimations of the transmitted codeword signals $\hat{\mathcal{S}}_t$. Upon successful decoding of packets, the respective transmitted codeword signals can be accurately reconstructed, facilitating interference cancellation to improve the performance of the JADD module. The iteration between the JADD module and the decoder bank stops when no more packets can be decoded, at which point the window advances by a step of Δ . For clarity, the receiver structure is described in Algorithm 1.

Algorithm 1. *The sliding-window receiver framework.*

- 1: Initialize t
 - 2: **repeat**
 - 3: $\mathbf{Y} = \mathbf{Y}^{(t)} \triangleq [\mathbf{y}_{t-L+1}, \dots, \mathbf{y}_t]$
 - 4: JADD: run the JADD algorithm and yield the estimations $\hat{\mathcal{A}}_t$, $\hat{\mathcal{D}}_t$, and $\hat{\mathcal{S}}_t$ based on $\mathbf{Y}$
 - 5: $\hat{\mathcal{A}}_t^D = \emptyset$, $\hat{\mathbf{X}}_{IC} = \mathbf{0}_{K \times L}$
 - 6: **for all** $k \in \hat{\mathcal{A}}_t$ **do**
-

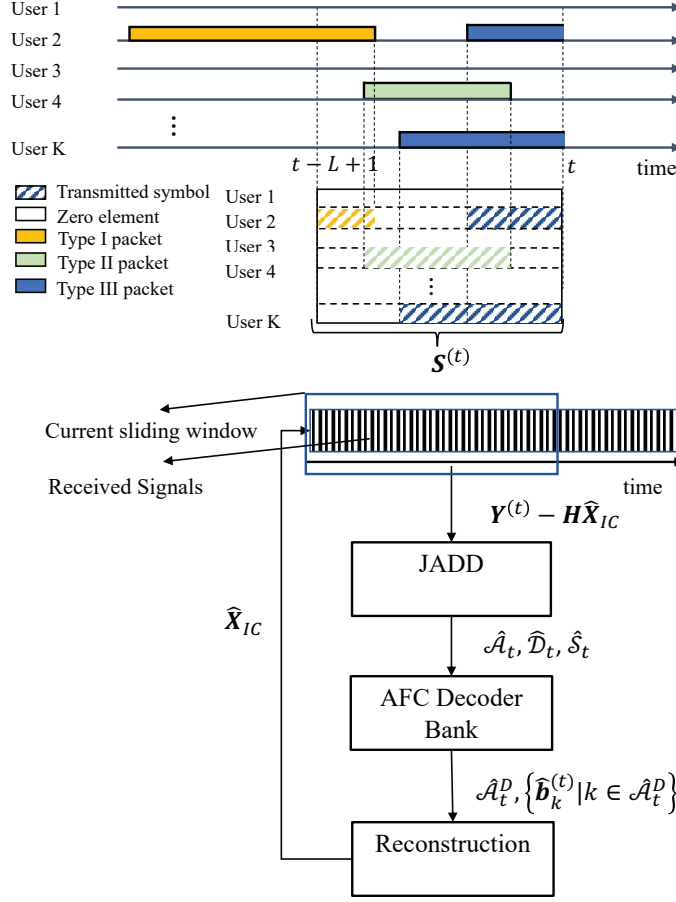


Figure 3. Illustration of the sliding-window framework and the receiver structure, where $\mathbf{S}^{(t)}$ represents the transmitted signal matrix of the current window.

- 7: AFC Decoding: run the AFC decoding algorithm based on the estimated codeword $\hat{\mathbf{c}}_k^{(t)} = [\hat{s}_{k,\hat{\tau}_k^{(t)}}^{(t)}, \dots, \hat{s}_{k,L}^{(t)}]$ to obtain the decoded message $\hat{\mathbf{b}}_k^{(t)}$
- 8: **if** user k is decoded **then**
- 9: $\hat{\mathcal{A}}_t^D = \hat{\mathcal{A}}_t^D \cup \{k\}$
- 10: Regenerate the transmitted codeword of user k : $\hat{\mathbf{X}}_{IC}[k, \hat{\tau}_k^{(t)} : L] = \Phi(\mathbf{G}_k \hat{\mathbf{b}}_k^{(t)})^T$
- 11: **end if**
- 12: **end for**
- 13: **until** $\hat{\mathcal{A}}_t^D = \emptyset$
- 14: Interference cancellation: $\mathbf{Y}^{(t)} = \mathbf{Y}^{(t)} - \mathbf{H}\hat{\mathbf{X}}_{IC}$
- 15: Window sliding: $t = t + \Delta$
- 16: Go back to step 2

III. FORMULATION OF THE JADD PROBLEM IN FA-GF-RMA

As mentioned in the previous section, the objective of JADD in FA-GF-RMA is to identify the users with an undecoded packet that initiates transmission within the sliding window, determine their transmission start times, and estimate their transmitted codeword signals. This section is dedicated to formulating the JADD problem in FA-GF-RMA.

Given that the transmitted signals from the decoded packets are effectively canceled, within a sliding window of length L (equal to the maximum transmission time), only the signals of the type III packets and type I packets are received. In other words, for each active user, a maximum of two undecoded packets from the same user can be observed within a given sliding

window. The first packet is of type I and represents an outage, while the second is of type III and remains undecoded.

The above observation can be formulated as follows. Let the relative ending point of the type I packet and the relative starting point of the type III packet for user k in the t -th sliding window be denoted as $\zeta_k^{(t)}$ and $\tau_k^{(t)}$, respectively. We define the tuple $(\zeta_k^{(t)}, \tau_k^{(t)})$ as the *activity mode* of user k in sliding window t . For all k and t , $(\zeta_k^{(t)}, \tau_k^{(t)}) \in \mathcal{M}$, where $\mathcal{M}$ is defined as $\{(\zeta, \tau) | 1 \leq \zeta < \tau \leq L\} \cup \{(0, \tau) | 1 \leq \tau \leq L\} \cup \{(\zeta, L+1) | 1 \leq \zeta < L\} \cup \{(0, L+1)\}$ and $|\mathcal{M}| = 3L^2/2 - L/2$.

It's worth noting that $\zeta_k^{(t)} = 0$ and $\tau_k^{(t)} = L+1$ respectively signify that no type I and type III packet from user k is transmitted in the t -th window, and $(\zeta_k^{(t)}, \tau_k^{(t)}) = (0, L+1)$ represents that user k is not active in the window.

For the convenience of description, we refer to $(0, L+1)$ as the inactive mode and the others as active modes. And an integer $m_k^{(t)} \in \{0, 1, \dots, |\mathcal{M}| - 1\}$ is used to denote the activity mode of user k in window t . Then the original activity mode is retrieved via two pre-defined one-to-one mappings f_ζ and f_τ as $(\zeta_k^{(t)}, \tau_k^{(t)}) = (f_\zeta(m_k^{(t)}), f_\tau(m_k^{(t)}))$. In particular, we set $f_\zeta(0) = 0$ and $f_\tau(0) = L+1$, indicating that $m_k = 0$ represents the inactive mode.

With the above notations, the set of active users of type III packets $\mathcal{A}_t$, the set of corresponding starting points $\mathcal{D}_t$, and the set of transmitted codeword signals $\mathcal{S}_t$ are formulated as follows,

$$\mathcal{A}_t = \{k | f_\tau(m_k^{(t)}) < L+1\}, \quad (7)$$

$$\mathcal{D}_t = \{\tau_k^{(t)} = f_\tau(m_k^{(t)}) | k \in \mathcal{A}_t\}, \quad (8)$$

$$\mathcal{S}_t = \{[s_{k,1}^{(t)}, \dots, s_{k,L}^{(t)}] | k \in \mathcal{A}_t\}. \quad (9)$$

Consider an arbitrary window and drop the superscript (t) for simplicity, the signal model is rewritten as

$$\mathbf{Y} = \mathbf{H}\mathbf{S} + \mathbf{N}. \quad (10)$$

In order to provide the estimations of $\mathcal{A}_t$, $\mathcal{D}_t$ and $\mathcal{S}_t$, it is necessary to recover the signal matrix $\mathbf{S}$ and the latent activity modes $\mathbf{m} = [m_1, \dots, m_K]^T$ based on the observations $\mathbf{Y}$.

The minimum mean square error (MMSE) estimator of $s_{k,l}$ is the posterior mean, which can be expressed

as

$$\hat{s}_{k,l} = \int s_{k,l} p(s_{k,l} | \mathbf{Y}) ds_{k,l} = \int s_{k,l} p(\mathbf{S} | \mathbf{Y}) d\mathbf{S}_{\setminus k,l}, \quad (11)$$

where $\mathbf{S}_{\setminus k,l}$ denotes the collection of all elements in $\mathbf{S}$ except $s_{k,l}$, and

$$\begin{aligned} p(\mathbf{S} | \mathbf{Y}) &= \frac{p(\mathbf{Y} | \mathbf{S}) p(\mathbf{S})}{p(\mathbf{Y})} \\ &= \frac{1}{\tilde{Z}} \prod_{g=1}^G \prod_{l=1}^L p(y_{g,l} | \mathbf{S}) \prod_{k=1}^K \sum_{m_k} p(\mathbf{S}_{k,:} | m_k) p(m_k), \end{aligned} \quad (12)$$

where $\tilde{Z}$ is the normalization constant.

The factorization of $p(\mathbf{Y} | \mathbf{S})$ and $p(\mathbf{S})$ in Eq. (12) is due to the assumption of AWGN and the independence across the users, respectively.

To provide the maximum a posteriori (MAP) estimation of m_k , the posterior distribution of m_k needs to be evaluated, which is expressed as

$$\hat{m}_k = \max_{m_k} \int p(\mathbf{m} | \mathbf{Y}) d\mathbf{m}_{\setminus k}, \quad (13)$$

where $\mathbf{m}_{\setminus k}$ denotes the collection of all elements in $\mathbf{m}$ except m_k , and

$$\begin{aligned} p(\mathbf{m} | \mathbf{Y}) &= \frac{\int p(\mathbf{m}, \mathbf{S}, \mathbf{Y}) d\mathbf{S}}{p(\mathbf{Y})} \\ &\propto p(\mathbf{m}) \int p(\mathbf{S} | \mathbf{m}) p(\mathbf{Y} | \mathbf{S}) d\mathbf{S} \\ &= \prod_{k=1}^K p(m_k) \int \prod_{k=1}^K p(\mathbf{S}_{k,:} | m_k) p(\mathbf{Y} | \mathbf{S}) d\mathbf{S}. \end{aligned} \quad (14)$$

IV. THE PROPOSED JADD ALGORITHM

Both the MMSE and MAP estimators necessitate the evaluation of posterior distributions that involve intractable integrals. Therefore, in this section, we propose an approximate approach to obtain the MMSE and MAP estimations.

First, exploiting the clustered sparsity structure of $\mathbf{S}$, the approximate message passing with nearest neighbor sparsity pattern learning (AMP-NNSPL) algorithm is employed to obtain $\hat{\mathbf{S}}$ and identify active users. Then, a power-pattern-based maximum

a posteriori (MAP) algorithm is introduced to estimate m_k for all detected active users, given by $\hat{m}_k = \arg \max_{m_k} p(m_k | \hat{\mathbf{S}}_{k,:})$. Finally, we propose a novel iterative scheme that combines the AMP-NNSPL and power-pattern-based MAP algorithms to enhance estimation performance.

4.1 The AMP-NNSPL Algorithm

Because of the sporadic nature of MMTC, only a small fraction of potential users are active at any given time. Consequently, the columns of $\mathbf{S}$ are sparse vectors. Therefore, the recovery of each column of $\mathbf{S}$ can be conceptualized as a single measurement vector (SMV) problem in the field of compressed sensing, which can be efficiently addressed using the AMP algorithm.

Specifically, considering the l -th symbol time within the sliding window, the received signal at this moment can be represented as

$$\mathbf{y}_l = \mathbf{H}\mathbf{s}_l + \mathbf{n}_l, \quad (15)$$

where $\mathbf{H}$ is the known channel matrix, and $\mathbf{s}_l = [s_{1,l}, \dots, s_{K,l}]^T$ is the sparse vector to be estimated.

To characterize the sparsity of $\mathbf{s}_l$, a flexible spike and slab prior distribution is adopted, i.e.,

$$p(\mathbf{s}_l) = \prod_{k=1}^K (1 - \gamma_{k,l}) \delta(s_{k,l}) + \gamma_{k,l} f(s_{k,l}), \quad (16)$$

where $f(s_{k,l})$ characterize the distribution of the l -th AFC-coded symbol of user k 's packets. In this work, we assume a Gaussian distribution for the AFC-coded symbols². Therefore, in Eq. (16), for all k and l , $f(s_{k,l}) = \mathcal{CN}(s_{k,l}; 0, \psi^2)$ with the variance ψ^2 evaluated as $\sum_{k,l} v_{k,l}^2 / K/L$.

At the q -th iteration, the AMP algorithm provides an estimation of $p(s_{k,l} | \mathbf{y}_l)$ as

$$\hat{p}(s_{k,l} | \mathbf{y}_l) \propto p(s_{k,l}) \mathcal{CN}(s_{k,l}; C_{k,l}^q, D_{k,l}^q), \quad (17)$$

which can be interpreted as the posterior distribution of $s_{k,l}$ given an observation $C_{k,l}^q = s_{k,l} + \bar{n}_{k,l}^q$, where $\bar{n}_{k,l}$ is the independent effective Gaussian noise with zero mean and variance $D_{k,l}^q$. Given the a priori model

in Eq. (16), the estimated posterior distribution is rearranged as

$$\begin{aligned} \hat{p}(s_{k,l} | \mathbf{y}_l) &= (1 - \pi_{k,l}^q) \delta(s_{k,l}) \\ &+ \pi_{k,l}^q \mathcal{CN}(s_{k,l}; A_{k,l}^q, B_{k,l}^q). \end{aligned} \quad (18)$$

It is worth noting that the estimated posterior distribution is a Bernoulli Gaussian distribution, in which $\pi_{k,l}^q$ can be interpreted as the posterior probability of $s_{k,l}$ being non-zero or belief indicator.

Assuming that the columns of $\mathbf{S}$ are mutually independent, the posterior distribution $p(s_{k,l} | \mathbf{Y}) = p(s_{k,l} | \mathbf{y}_l)$. Then the posterior expectation $\bar{s}_{k,l}^q$ and variance $\nu_{k,l}^q$ at the q -th iteration are respectively given as

$$\bar{s}_{k,l}^q = \int s_{k,l} \hat{p}(s_{k,l} | \mathbf{y}_l) ds_{k,l} = \pi_{k,l}^q A_{k,l}^q, \quad (19)$$

$$\begin{aligned} \nu_{k,l}^q &= \int |s_{k,l} - \bar{s}_{k,l}^q|^2 \hat{p}(s_{k,l} | \mathbf{y}_l) ds_{k,l} \\ &= \pi_{k,l}^q (|A_{k,l}^q|^2 + B_{k,l}^q) - |\bar{s}_{k,l}^q|^2. \end{aligned} \quad (20)$$

And the estimation of $s_{k,l}$ is provided as $\hat{s}_{k,l} = \bar{s}_{k,l}$.

Nevertheless, independently recovering L columns of $\mathbf{S}$ via AMP does not leverage the inherent sparsity structure and may consequently result in unsatisfactory performance. As depicted in Figure 3, the non-all-zero rows of $\mathbf{S}$ exhibit a clustered sparsity structure, where the non-zero elements are grouped into a block. Similar clustered sparsity structures have appeared in the previous literature [38, 39], where the nearest neighbor sparsity pattern learning (NNSPL) technique is applied to the original AMP algorithm. The rationale behind NNSPL is that $s_{k,l}$ and its neighboring elements, such as $s_{k,l-1}$ and $s_{k,l+1}$, tend to be either simultaneously zero or non-zero.

In this paper, we adopt the AMP-NNSPL algorithm to accommodate the clustered sparsity structure exhibited in FA-GF-RMA. In Algorithm 2, $\mathcal{N}_l \triangleq \{l-1, l+1\}$ represents the two neighbors of the l -th element and q_{max} is the predefined number of iterations.

The AMP-NNSPL algorithm directly provides the estimation of transmitted signals $\hat{\mathbf{S}}_t$. To obtain $\hat{\mathbf{A}}_t$, activity detection is required. A method of activity detection based on the belief indicators is developed in [39]. For convenience of description, we define a

symbol-level activity state $\alpha_{k,l}$ as

$$\alpha_{k,l} = \begin{cases} 1, & \text{if } f_\tau(m_k) \leq l \text{ or } f_\zeta(m_k) \geq l, \\ 0, & \text{otherwise,} \end{cases} \quad (21)$$

and define a threshold function $r(x; \epsilon)$, where $r(x; \epsilon) = 1$ if $|x| > \epsilon$ and otherwise $r(x; \epsilon) = 0$.

The activity detection begins by making a threshold decision on symbol-level activity states, which employs

$$\hat{\alpha}_{k,l} = r(\pi_{k,l}; \epsilon_{bi}). \quad (22)$$

Subsequently, the user-level activity is determined as

$$\hat{\alpha}_k = r\left(\frac{1}{L} \sum_l \hat{\alpha}_{k,l}, p_{th}\right). \quad (23)$$

Then the active user set is estimated as $\hat{\mathcal{A}}_t = \{k | \hat{\alpha}_k = 1\}$. Note that the threshold p_{th} can be considered the required proportion of active symbols to categorize a user as active. In general, a higher p_{th} results in a set that includes only users whose transmitted codewords are sufficiently lengthy, whereas lowering p_{th} reduces the risk of miss detection but raises the false alarm rate.

However, activity detection can only identify whether a user is active but not able to distinguish among active modes. In the next sub-section, we develop a power-pattern-based activity mode estimation algorithm.

Algorithm 2. AMP-NNSPL.

Input: The received signal $\mathbf{Y}$, the prior active probability matrix $\mathbf{\Gamma}$, the sensing matrix $\mathbf{H}$

- 1: Initialization: $\forall k, l, \nu_{k,l}^0 = \psi^2, \hat{s}_{k,l}^0 = 0, z_{g,l}^0 = y_{g,l}, q = 1$
- 2: **repeat**
- 3: $\forall g, l :$
- 4: $\kappa_{g,l}^q = \sum_k |H_{g,k}|^2 \nu_{k,l}^{q-1}$
- 5: $z_{g,l}^q = \sum_k H_{g,k} \hat{s}_{k,l}^q - \frac{\kappa_{g,l}^q}{\sigma^2 + \kappa_{g,l}^{q-1}} (y_{g,l} - z_{g,l}^{q-1})$
- 6: $\forall k, l :$
- 7: $D_{k,l}^q = \left[\sum_g \frac{|H_{g,k}|^2}{\sigma^2 + \kappa_{g,l}^q} \right]^{-1}$
- 8: $C_{k,l}^q = \hat{s}_{k,l}^q + D_{k,l}^q \sum_g \frac{h_{g,k}^* (y_{g,l} - z_{g,l}^q)}{\sigma^2 + \kappa_{g,l}^q}$
- 9: $\chi_{k,l}^q = \ln \frac{D_{k,l}^q}{D_{k,l}^q + \psi^2} + \frac{|C_{k,l}^q|^2}{D_{k,l}^q} - \frac{|C_{k,l}^q|^2}{D_{k,l}^q + \psi^2}$
- 10: $\pi_{k,l}^q = \frac{\gamma_{k,l}}{\gamma_{k,l} + (1 - \gamma_{k,l}) \exp(-\chi_{k,l}^q)}$

- 11: $A_{k,l}^q = \frac{\psi^2 C_{k,l}^q}{D_{k,l}^q + \psi^2} \quad B_{k,l}^q = \frac{\psi^2 D_{k,l}^q}{D_{k,l}^q + \psi^2}$
- 12: $\hat{s}_{k,l}^q = \pi_{k,l}^q A_{k,l}^q$
- 13: $\nu_{k,l}^q = \pi_{k,l}^q (|A_{k,l}^q|^2 + B_{k,l}^q) - |\hat{s}_{k,l}^q|^2$
- 14: $\gamma_{k,l} = \frac{1}{|\mathcal{N}_l|} \sum_{j \in \mathcal{N}_l} \pi_{k,j}^q$
- 15: $q = q + 1$
- 16: **until** $q > q_{max}$

Output: The belief indicator matrix $\mathbf{\Pi}$ where $\pi_{k,l} = \pi_{k,l}^q$, and the estimated signal matrix $\hat{\mathbf{S}}$

4.2 Power-Pattern-Based Activity Mode Estimation

As mentioned in the previous sub-section, the AMP-NNSPL algorithm falls short in estimating the activity mode m_k , making it incapable of providing estimates for the starting points of type III packets. Despite its ability to produce belief indicators for symbol-level activity, they lack the required reliability to determine the starting point, as depicted in Figure 4.

Moreover, in cases where a codeword begins with a zero symbol, which is possible with the AFC code, the related belief indicator typically approaches nearly zero. This situation will result in an erroneous identification of the packet's starting point via the belief indicators, delayed by one symbol.

To address the activity mode estimation problem, we resort to the estimation of transmitted signals. From Eq. (4), it is evident that the transmission power of AFC-coded symbols fluctuates over time due to the variable degree or power gain factor. This power variation, referred to as the power pattern, can be harnessed. By purposefully designing the power mode to exhibit favorable auto-correlation properties, noticeable second-order statistical distinctions can emerge among the transmitted signals of packets starting at different time instances. These distinctions can subsequently be employed to distinguish the starting points of the packets.

Notice that given the activity mode m_k , the elements in $\mathbf{S}_{k,:}$ are mutually independent. Therefore, $p(\mathbf{S}_{k,:} | m_k) = \prod_l p(s_{k,l} | m_k)$, where

$$p(s_{k,l} | m_k) = \begin{cases} f_{k,L-f_\zeta(m_k)+1}(s_{k,l}), & 1 \leq l \leq f_\zeta(m_k), \\ f_{k,l-f_\tau(m_k)+1}(s_{k,l}), & f_\tau(m_k) \leq l \leq L, \\ \delta(s_{k,l}), & \text{otherwise.} \end{cases} \quad (24)$$

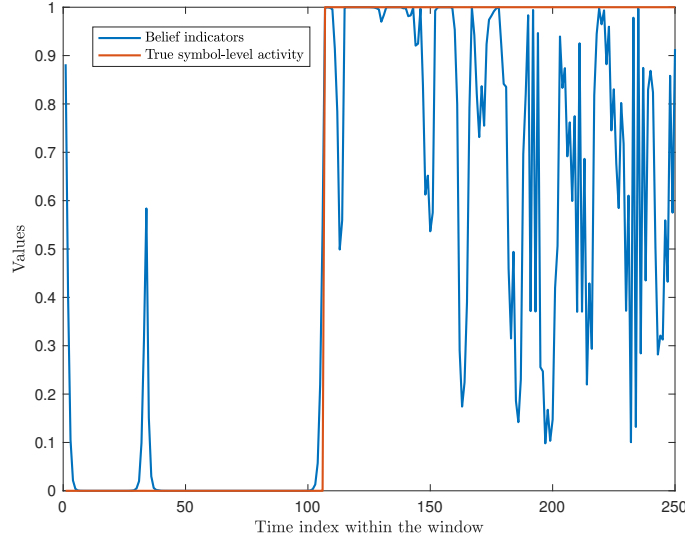


Figure 4. The typical per-symbol posterior probability curves of active users output by the AMP-SSL algorithm are shown with $K = 500$, $M = 50$, and $\lambda = 0.0002$.

Here $f_{k,l}(\cdot)$ denotes the distribution of the l -th transmitted AFC-coded symbol of user k 's packets. With the Gaussian assumption of AFC-coded symbols, $f_{k,l}(\cdot) = \mathcal{CN}(\cdot; 0, v_{k,l}^2)$, where $v_{k,l}^2$ is given by Eq. (4).

Therefore, given the activity mode m_k , the transmitted signals within a sliding window $[s_{k,l}, \dots, s_{k,L}]$ can be regarded as a zero-mean, variance-varying discrete Gaussian process, where the conditional variance $\bar{v}_{m_k,l}^2$ of $s_{k,l}$ is given by

$$\bar{v}_{m_k,l}^2 = \begin{cases} v_{k,L-f_\zeta(m_k)+l}^2, & 1 \leq l \leq f_\zeta(m_k), \\ v_{k,l-f_\tau(m_k)+1}^2, & f_\tau(m_k) \leq l \leq L, \\ 0, & \text{otherwise.} \end{cases} \quad (25)$$

Now decompose the estimated signals as

$$\hat{s}_{k,l} = s_{k,l} + e_{k,l}, \quad \forall k, l, \quad (26)$$

where the estimation error $e_{k,l}$ is modeled as a Gaussian noise with distribution $\mathcal{CN}(0, \phi_{k,l}^2)$.

Then the posterior probability of m_k given the estimated signals $\hat{\mathbf{S}}_{k,:}$ is expressed as

$$p(m_k | \hat{\mathbf{S}}_{k,:}) \propto p(\hat{\mathbf{S}}_{k,:} | m_k) p(m_k)$$

$$= \prod_{l=1}^L \frac{1}{\pi(\bar{v}_{m_k,l}^2 + \phi_{k,l}^2)} \exp \left\{ -\frac{|\hat{s}_{k,l}|^2}{\bar{v}_{m_k,l}^2 + \phi_{k,l}^2} \right\} p(m_k). \quad (27)$$

According to Eq. (19), $\hat{s}_{k,l}$ can be viewed as an MMSE estimator, then the estimation error variance $\phi_{k,l}^2$ is given by the expectation of the posterior variance $\nu_{k,l}$ in Eq. (20), of which the closed-form expression is hard to obtain. In this study, we set $\phi_{k,l}^2$ equal to the posterior variance $\nu_{k,l}$.

Then the MAP estimation $\hat{m}_k$ is obtained as

$$\hat{m}_k = \arg \max_{m_k} \ln p(m_k | \hat{\mathbf{S}}_{k,:}), \quad (28)$$

and the estimated starting point is $f_\tau(\hat{m}_k)$.

Note that the activity mode estimation is carried out on a per-user basis, which can be computationally intensive, especially when dealing with a large number of potential users. To reduce complexity, it is feasible to conduct the estimation solely for users within the set $\hat{\mathcal{A}}_t = \{k | \hat{\alpha}_k = 1\}$, which is obtained through the activity detection process.

4.3 The Proposed Power Pattern Aided AMP Algorithm

A straightforward approach for JADD in FA-GF-RMA is to concatenate the power-pattern-based activity mode estimation module after the AMP-NNSPL algorithm. However, since prior information on power patterns is utilized during the activity mode estimation phase, it could be advantageous to iterate between these two steps to fully leverage the power patterns and enhance both algorithms' performance.

Notice that Eq. (27) provides a posterior distribution of the activity mode given the estimated signals $\hat{\mathbf{S}}_{k,:}$. Based on the relation between the activity mode and the symbol-level activity states as shown in Eq. (21), it is feasible to derive a posterior distribution of the symbol-level activity state $\alpha_{k,l}$ as

$$\hat{a}_{k,l} \triangleq p(\alpha_{k,l} = 1 | \hat{\mathbf{S}}_{k,:}) = \sum_{m_k \in \mathcal{M}_l} p\{m_k | \hat{\mathbf{S}}_{k,:}\}, \quad (29)$$

where $\mathcal{M}_l = \{m | \forall m \in \mathcal{M}, f_\tau(m) \leq l \text{ or } f_\zeta(m) \geq l\}$.

Intuitively, the activity mode estimation accuracy improves with the number of transmitted coded symbols in a window, as different modes exhibit greater distinctions in power patterns over a longer range of transmitted codewords. Consequently, the reliability of the posterior symbol-level active probability $\hat{a}_{k,l}$ improves with the codeword length. Therefore, the power-pattern information from longer-codeword users can be applied to the AMP-NNSPL algorithm by updating the corresponding symbol-level active probabilities $\gamma_{k,l}$ as $\hat{a}_{k,l}$, potentially enhancing the estimation quality of the signal of the shorter-codeword users. This process can be iteratively executed to fully leverage the power-pattern information from all active users in descending order of codeword length.

The above intuition is implemented in the proposed power-pattern-aided AMP (PP-AMP) algorithm 3, which can be summarized as the following 3 steps

- Step 1: Run the AMP-NNSPL to obtain the estimated signal matrix $\hat{\mathbf{S}}$ and the belief indicators. (Line 3)
- Step 2: Perform the active user detection with a decreasing threshold p_{th} until enough active users are detected. (Line 4-13)
- Step 3: Run the power-pattern-based activity

mode estimation procedure for the detected active users and update their prior symbol-level active probabilities. (Line 15-21)

Notice that the threshold p_{th} decreases from p_{IN} to p_0 with a step p_δ during the iteration. A higher p_{IN} ensures that users with sufficiently long codewords are initially detected. As p_0 is increased, users with relatively short codewords are excluded from the iteration, as their power-pattern information might be unreliable and unhelpful.

Algorithm 3. Power-pattern-aided AMP.

Input: The received signal $\mathbf{Y}$, the prior active probability matrix $\mathbf{\Gamma}_0$, the sensing matrix $\mathbf{H}$, initial activity detection threshold p_{IN} , and I_{Turbo} , I_{AMP} , p_δ , p_0

- 1: Initialization: $i = 0$, $\mathbf{\Gamma} = \mathbf{\Gamma}_0$, $\hat{\mathcal{A}} = \hat{\mathcal{A}}_{pre} = \emptyset$, $p_{th} = p_{IN}$
- 2: **repeat**
- 3: Run AMP-NNSPL algorithm for inputs $\mathbf{\Gamma}$, $\mathbf{Y}$, $\mathbf{H}$ and output $\hat{\mathbf{S}}$, $\mathbf{\Pi}$
- 4: Active user detection:
- 5: $\forall k, l, \hat{\alpha}_{k,l} = r(\pi_{k,l}; \epsilon_{bi})$
- 6: **repeat**
- 7: $\forall k, \hat{\alpha}_k = r(\frac{1}{L} \sum_l \hat{\alpha}_{k,l}, p_{th})$
- 8: $\hat{\mathcal{A}} = \{k | \hat{\alpha}_k = 1\}$
- 9: $p_{th} = \max\{p_{th} - p_\delta, p_0\}$
- 10: **if** $p_{th} = p_0$ **then**
- 11: break
- 12: **end if**
- 13: **until** $|\hat{\mathcal{A}}| > |\hat{\mathcal{A}}_{pre}|$
- 14: $\hat{\mathcal{A}}_{pre} = \hat{\mathcal{A}}$
- 15: Power-pattern-based activity mode estimation:
- 16: **if** $\hat{\mathcal{A}} \neq \emptyset$ **then**
- 17: $\forall k \in \hat{\mathcal{A}}, m_k \in \mathcal{M}$:
Compute $p\{m_k | \hat{\mathbf{S}}_{k,:}\}$ according to Eq. (27) and normalize it to a probability distribution
- 18: $\hat{m}_k = \max_{m_k \in \mathcal{M}} p\{m_k | \hat{\mathbf{S}}_{k,:}\}$
- 19: $\mathbf{\Gamma} = \mathbf{\Gamma}_0$
- 20: Update the symbol-level prior active probability:
- 21: $\forall k \in \hat{\mathcal{A}}, 1 \leq l \leq L$:
Compute $\hat{a}_{k,l}$ according to Eq. (29)
- 22: $\mathbf{\Gamma}_{k,l} = \hat{a}_{k,l}$
- 22: **end if**
- 23: $i = i + 1$
- 24: $p_{th} = p_{th} - p_\delta$
- 25: **if** $p_{th} \leq 0$ **then**

```

26:     break
27:   end if
28: until  $i > I_{Turbo}$ 

```

Output: Estimated signal matrix $\hat{\mathbf{S}}$, estimated active user set $\hat{\mathcal{A}}$, estimated activity mode $\{\hat{m}_k | k \in \hat{\mathcal{A}}\}$

4.4 Computational Complexity Analysis

The computational complexity of the algorithms is quantified in terms of the number of complex-valued scalar multiplications. In the AMP-NNSPL algorithm, each iteration requires approximately $7GKL + 2GL + 15KL$ multiplications. In the power-pattern-based activity mode estimation, evaluating a posterior distribution of $|\mathcal{M}|$ values is necessary for each active user, with each probability value involving $3L$ multiplications, and therefore $3K_aL|\mathcal{M}| = \frac{3}{2}K_a(L^3 - L^2)$ multiplications are required in each iteration. Here K_a represents the number of active users, which is much less than the number of potential users K .

Significant computational complexity arises from the large number of activity modes $\mathcal{M}$, which scales on the order of L^2 . In practice, given the infrequent occurrence of outage events, it is possible to work with a reduced set of activity modes that excludes those related to type I packets, i.e., $\mathcal{M} = \{(0, \tau) | 1 \leq \tau \leq L + 1\}$. In this case, the complexity of activity mode estimation is reduced to $3K_a(L^2 + L)$ per iteration.

V. SIMULATION RESULTS

In this section, Monte Carlo simulations are conducted to verify the performance of the proposed JADD algorithm in FA-GF-RMA. In the simulations, the number of users is set to $K = 500$, and the packet size is fixed at 100 bits. Each user generates a data packet flow according to a Poisson process with an arrival rate of λ per symbol interval. The large-scale coefficients of all users are normalized, i.e., $\beta_k = 1$. The signal-to-noise ratio (SNR) is defined as the ratio of average transmitting power to noise power. The information bits of user data packets are not pre-coded and are directly input to the AFC encoder after BPSK modulation. A fixed-degree AFC code with a degree of 8 and a weight set $\{1, 2, 3, 4\}$ is employed [20], and its energy is normalized³. For power pattern generation, the power gain factor $p_{k,l}$ in Eq. (3) is independently and uniformly sampled from the set $\{0.4, 1.6\}$. The AFC code

has a maximum length of $L = 250$, and AFC decoding is accomplished using the BP algorithm. The receiver attempts decoding whenever the window moves forward with a step size of $\Delta = 2$. Successful decoding is confirmed when the BPSK sequence of the data packet is fully restored.

First, we run simulations according to the signal model Eq. (10). In these simulations, to mimic the sparsity structure of $\mathbf{S}$, each user is activated with a probability of K_a/K , where K_a implies the average number of active users. Each active user then selects a starting point drawn from either the range $[131, 250]$ with a probability of 0.8 or $[51, 130]$ with a probability of 0.2. For performance evaluation, we use the metrics of mean squared error (MSE), normalized mean squared error of codeword (NMSE-C), and starting point estimation error rate (SPER), which are defined respectively as

$$\text{MSE} = \frac{\|\hat{\mathbf{S}} - \mathbf{S}\|_F^2}{KL}, \quad (30)$$

$$\text{NMSE-C} = \frac{\|\hat{\mathbf{S}}_{k,L-l+1:L} - \mathbf{S}_{k,L-l+1:L}\|^2}{\|\mathbf{S}_{k,L-l+1:L}\|^2}, \quad (31)$$

$$\text{SPER} = \frac{\sum_{k \in \{k | m_k \neq 0\}} \mathbb{I}\{\hat{m}_k \neq m_k\}}{\sum_k \mathbb{I}\{m_k \neq 0\}}, \quad (32)$$

$$(33)$$

where l represents the codeword length and $\mathbb{I}\{\cdot\}$ is the indicator function.

Figure 5 presents a comparison of the MSE performance between the AMP-NNSPL algorithm and the proposed PP-AMP algorithm. The benchmark used for this comparison is the Oracle Linear Minimum Mean Squared Error (LMMSE) algorithm, which assumes knowledge of symbol-level activities and recovers the signal matrix column by column through LMMSE. The results clearly demonstrate that the proposed PP-AMP outperforms the original AMP-NNSPL algorithm, closely approaching the performance of the Oracle LMMSE. The MSE reduction of PP-AMP concerning AMP-NNSPL becomes more significant as the number of antennas decreases or the average number of active users increases. This suggests that the proposed PP-AMP algorithm effectively leverages the sparsity structure, enabling support for more active users with fewer antennas.

Figure 6 and Figure 7 compare the NMSE-C and

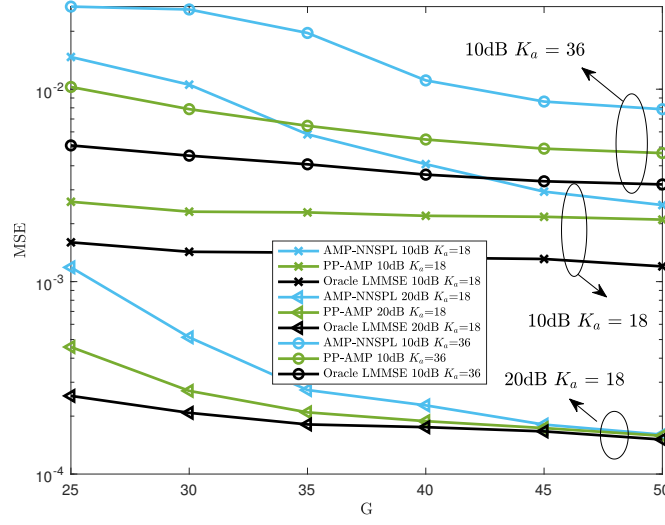


Figure 5. Comparison of MSE versus the number of antennas for the proposed PP-AMP and the state-of-art AMP-NNSPL algorithms.

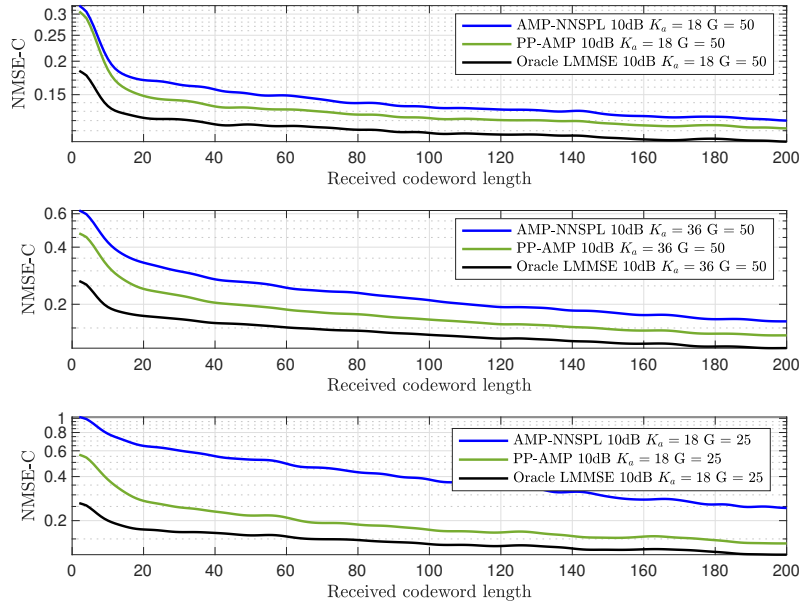


Figure 6. Comparison of NMSE-c versus the length of received codeword length l for the proposed PP-AMP and the state-of-art AMP-NNSPL algorithms.

SPER performance for the various algorithms as a function of received codeword length. For the AMP-NNSPL and Oracle LMMSE algorithms, the SPER curves are obtained by conducting the starting point estimation for the estimated signals.

The results clearly demonstrate a trend where both NMSE-C and SPER decrease as the received code-

word length increases, indicating that estimations for packets with longer codewords are more reliable. Notably, the NMSE-C and SPER of the PP-AMP algorithm exhibit a notably faster decrease, especially under conditions of fewer antennas or higher active probability. It is also worth noting that the NMSE-C and SPER of both AMP-NNSPL and PP-AMP algorithms

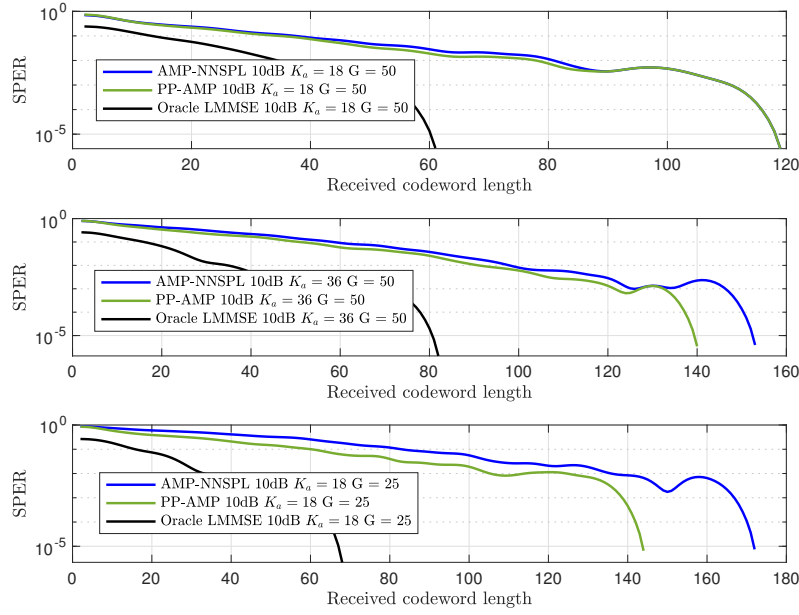


Figure 7. Comparison of SPER versus the length of received codeword length l for the proposed PP-AMP and the state-of-art AMP-NNSPL algorithms.

are relatively higher than that of Oracle LMMSE, especially at small received codeword lengths. This might be due to the approximations of the original estimation problems in Eq. (12) and Eq. (14). Jointly estimation of the signal matrix and the latent activity modes through factor-graph-based inference techniques might alleviate this problem, which will be studied in future work.

Next, we proceed to compare the delay performance of both FA-GF-RMA and FS-GF-RMA. The FS-GF-RMA scheme employed here is beacon-based, as described in [40], where the base station initiates transmission periods via downlink beacons. In this scheme, a user can only initiate access at the beginning of a transmission period. The receiver aims to decode all active users during a transmission period whenever additional Δ coded symbols are received. The transmission period concludes when all active users successfully decode their data or when the maximum codeword length is reached. In both schemes, the Oracle LMMSE algorithm provides the lower bound of delay achievable by any specific JADD algorithms. Notice that in FA-GF-RMA with Oracle LMMSE, the starting points are perfectly known. As for FA-GF-RMA with the AMP-NNSPL algorithm, the starting points

are estimated by conducting the power-pattern-based activity mode estimation procedure afterward.

The overall delay D_o encompasses the time duration from the moment a packet arrives until it is successfully decoded. The transmission delay D_t is defined as the time interval from the point of packet access to successful decoding. The difference between D_o and D_t is the waiting delay. It's worth noting that in FA-GF-RMA, each packet initiates access upon reaching the buffer, provided the buffer is empty. As a result, D_t is almost equivalent to D_o in FA-GF-RMA, i.e., no waiting delay incurred, especially given a relatively low packet arrival rate.

Figure 8 illustrates the average delays at different SNRs with an antenna number of $G = 50$ and a packet arrival rate of $\lambda = 0.0005$ per symbol interval per user. As can be observed, the delays of the AMP-NNSPL and PP-AMP algorithms are close to each other, and both algorithms approach the lower bound as SNRs increase. This is because these algorithms demonstrate similar estimation performance, particularly at high SNRs and antenna numbers, as depicted in Figure 5. It's essential to emphasize that, although the lower bound of transmission delay is smaller for FS-GF-RMA compared to FA-GF-RMA, the substantial

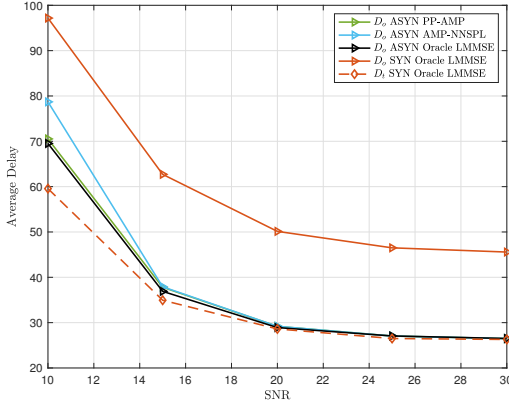


Figure 8. Comparison of average delays of FA-GF-RMA and FS-GF-RMA at different SNRs with $G = 50$ and $\lambda = 5 \times 10^{-4}$ packets per user per symbol interval.

waiting delay in FS-GF-RMA, which accounts for approximately half of the transmission delay, results in a significantly longer overall delay compared to FA-GF-RMA.

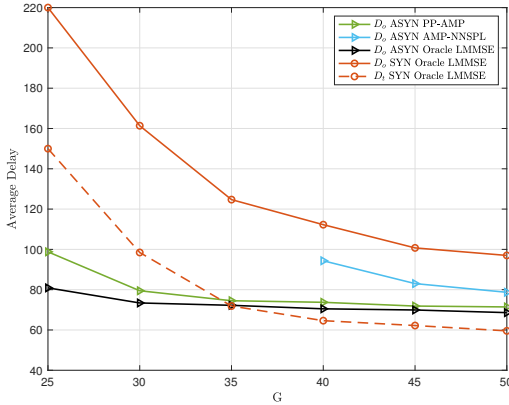


Figure 9. Comparison of average delays of FA-GF-RMA and FS-GF-RMA at different antenna numbers with $\text{SNR}=10\text{dB}$ and $\lambda = 5 \times 10^{-4}$ packets per user per symbol interval.

Figure 9 provides a comparison of delays about the number of antennas G . The proposed PP-AMP algorithm approaches the delay achieved by Oracle LMMSE. However, it's important to note that FA-GF-RMA with the AMP-NNSPL algorithm becomes unstable when the number of antennas is less than 40. As a result, delay values for these configurations are not displayed. This instability arises because the NMSE-C of the AMP-NNSPL algorithm decreases too slowly

when the number of antennas is small, as depicted in the third sub-graph in Figure 6. Consequently, the decoder fails to decode packets and perform interference cancellation at a sufficient rate before more newly arrived packets enter the observation window.

To illustrate this, a sample trajectory of the number of active users in the observation window at $G = 30$ is presented in Figure 10. As can be seen, with the AMP-NNSPL algorithm, the number of active users continues to increase after surpassing 25, ultimately exceeding the total number of antennas, which is 30. It's also worth noting that the count of active users in an observation window changes over time in FA-GF-RMA. When the system reaches a stable state, the average number of active users is roughly equivalent to $\lambda K D_t$, which is approximately $0.0005 \times 500 \times 80 = 20$ in the example presented in Figure 10.

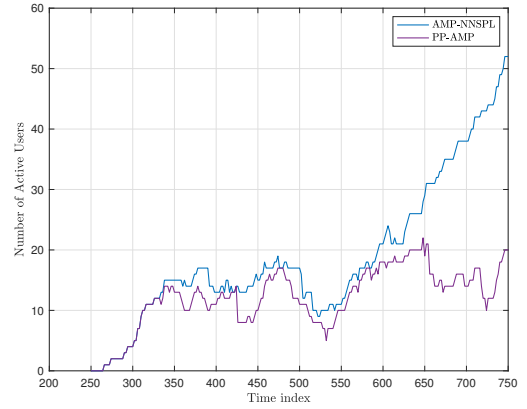


Figure 10. A sample trajectory of the number of active users versus time in FA-GF-RMA at $G = 30$, $\lambda = 5 \times 10^{-4}$ packets per symbol per user and $\text{SNR}=10\text{ dB}$.

Figure 11 illustrates the delays of both FA-GF-RMA and FS-GF-RMA using different algorithms versus the arrival rate. Similarly, the proposed PP-AMP algorithm exhibits delays close to the lower bound, while the AMP-NNSPL algorithm renders FA-GF-RMA unstable at high arrival rates. It's also noticeable that FS-GF-RMA is unstable, even with the Oracle LMMSE algorithm, at a high arrival rate of 0.0011 packets per symbol interval. The rationale behind this phenomenon is as follows: In FS-GF-RMA with Oracle LMMSE and a sufficient number of antennas G , the multi-user interference is nearly completely suppressed. Nevertheless, with an excessively high arrival rate, the number of newly arrived users during

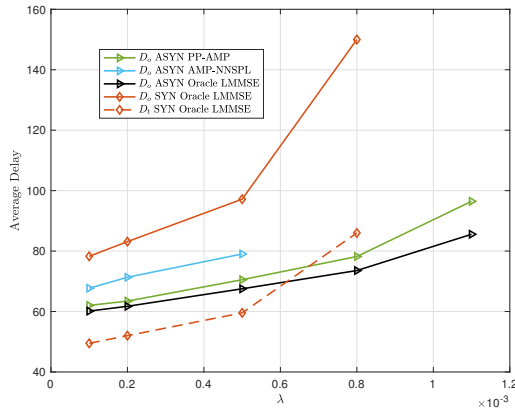


Figure 11. Comparison of average delays of FA-GF-RMA and FS-GF-RMA at different arrival rates with $G = 50$ and $SNR=10dB$.

the current transmission period can surpass G , resulting in the degradation of interference suppression during the following transmission period. Consequently, the transmission period extends and encompasses even more newly arrived users, ultimately leading to system divergence.

VI. CONCLUSION

In this work, a joint user activity and data detection algorithm is proposed for the FA-GF-RMA system. Considering the sparsity in user activities, an existing AMP-based algorithm for solving the JADD problem in synchronous grant-free access is directly applied. To address the limitation of the AMP algorithm in estimating the transmission starting point, the intrinsic power pattern in the rateless codeword is exploited to tackle this issue. Simulation results demonstrate that the proposed power-pattern-aided AMP algorithm outperforms the existing AMP-NNSPL algorithm in recovering the signal matrix with the unique sparsity structure observed in FA-GF-RMA. The delays achieved by the proposed solution in FA-GF-RMA with AFC approach the delay lower bound across various system configurations, verifying the proposed solution achieves the delay advantages of FA-GF-RMA over FS-GF-RMA in the practical scenario with unknown user activity information and practical rateless codes.

ACKNOWLEDGEMENT

This work is supported by the projects as follows, Key Research and Development Program of China (2018YFB1801102), Fundamental Research Funds for the Central Universities under Grant 2242022k60006, the Key Research and Development Program of China (2020YFB1806603), Tsinghua University-China Mobile Communications Group Co., Ltd. Joint Institute, Civil Aerospace Technology Project (D040202), National Natural Science Foundation of China (Grant No.92067206, Tsinghua-Qualcomm Joint Project, Tsinghua University Initiative Scientific Research Program (20193080005).

NOTES

¹It is assumed that the ACK is transmitted through an error-free channel with zero delays.

²Notice that an AFC-coded symbol is essentially the summation of independent random variables. According to the central limit theorem, when the degree is sufficiently large, the AFC-coded symbol approaches a Gaussian random variable.

³The optimization of pre-code, degree distribution and weight set are beyond the scope of this paper.

References

- [1] H. S. Dhillon, H. Huang, *et al.*, "Wide-area wireless communication challenges for the internet of things," *IEEE Communications Magazine*, 2017, vol. 55, no. 2, pp. 168–174.
- [2] Z. Dawy, W. Saad, *et al.*, "Toward massive machine type cellular communications," *IEEE Wireless Communications*, 2016, vol. 24, no. 1, pp. 120–128.
- [3] M. Vaezi, A. Azari, *et al.*, "Cellular, wide-area, and non-terrestrial iot: A survey on 5g advances and the road towards 6g," *IEEE Communications Surveys & Tutorials*, 2022.
- [4] C. Bockelmann, N. Pratas, *et al.*, "Massive machine-type communications in 5g: Physical and mac-layer solutions," *IEEE Communications Magazine*, 2016, vol. 54, no. 9, pp. 59–65.
- [5] M. B. Shahab, R. Abbas, *et al.*, "Grant-free non-orthogonal multiple access for iot: A survey," *IEEE Communications Surveys & Tutorials*, 2020, vol. 22, no. 3, pp. 1805–1838.
- [6] E. Dahlman, S. Parkvall, *et al.*, *3G Evolution: HSPA and LTE for Mobile Broadband*. Academic Press, 2010.
- [7] 3GPP, "Technical Specification Group Radio Access Network; NR; NR and NG-RAN Overall Description," 3rd Generation Partnership Project (3GPP), Technical Specification (TS) 38.300, version 16.0.0. 2019.

- [8] M. Hasan, E. Hossain, *et al.*, "Random access for machine-to-machine communication in lte-advanced networks: Issues and approaches," *IEEE Communications Magazine*, 2013, vol. 51, no. 6, pp. 86–93.
- [9] S. R. Pokhrel, J. Ding, *et al.*, "Towards enabling critical mmhc: A review of urllc within mmhc," *IEEE Access*, 2020, vol. 8, pp. 131 796–131 813.
- [10] L. Tian, C. Yan, *et al.*, "On uplink non-orthogonal multiple access for 5g: Opportunities and challenges," *China Communications*, 2017, vol. 14, no. 12, pp. 142–152.
- [11] F. Jabbarvaziri, N. M. Balasubramanya, *et al.*, "Harq-based grant-free noma for mmhc uplink," *IEEE Internet of Things Journal*, 2020, vol. 8, no. 10, pp. 8372–8386.
- [12] Z. Zhang, X. Wang, *et al.*, "Rateless multiple access: Asymptotic throughput analysis and improvement with spatial coupling," *IEEE Access*, 2018, vol. 6, pp. 63 200–63 213.
- [13] Y. Yuan, S. Wang, *et al.*, "Noma for next-generation massive iot: Performance potential and technology directions," *IEEE Communications Magazine*, 2021, vol. 59, no. 7, pp. 115–121.
- [14] M. Shirvanimoghaddam, M. S. Mohammadi, *et al.*, "Short block-length codes for ultra-reliable low latency communications," *IEEE Communications Magazine*, 2018, vol. 57, no. 2, pp. 130–137.
- [15] X. Chen and R. Jia, "Exploiting rateless coding for massive access," *IEEE Transactions on Vehicular Technology*, 2018, vol. 67, no. 11, pp. 11 253–11 257.
- [16] M. Shirvanimoghaddam, Y. Li, *et al.*, "Near-capacity adaptive analog fountain codes for wireless channels," *IEEE Communications Letters*, 2013, vol. 17, no. 12, pp. 2241–2244.
- [17] J. Perry, H. Balakrishnan, *et al.*, "Rateless spinal codes," in *Proceedings of the 10th ACM Workshop on Hot Topics in Networks*, 2011, pp. 1–6.
- [18] H. Chen, R. Abbas, *et al.*, "Ultra-reliable low latency cellular networks: Use cases, challenges and approaches," *IEEE Communications Magazine*, 2018, vol. 56, no. 12, pp. 119–125.
- [19] R. Abbas, M. Shirvanimoghaddam, *et al.*, "Novel design for short analog fountain codes," *IEEE Communications Letters*, 2019, vol. 23, no. 8, pp. 1306–1309.
- [20] K. Zhang, J. Jiao, *et al.*, "Finite block-length analog fountain codes for ultra-reliable low latency communications," *IEEE Transactions on Communications*, 2020, vol. 68, no. 3, pp. 1391–1404.
- [21] Y. Polyanskiy, H. V. Poor, *et al.*, "Channel coding rate in the finite blocklength regime," *IEEE Transactions on Information Theory*, 2010, vol. 56, no. 5, pp. 2307–2359.
- [22] M. Shirvanimoghaddam, M. Condoluci, *et al.*, "On the fundamental limits of random non-orthogonal multiple access in cellular massive iot," *IEEE Journal on Selected Areas in Communications*, 2017, vol. 35, no. 10, pp. 2238–2252.
- [23] W. Zhang, S. Zhou, *et al.*, "An asynchronous grant-free multiple access scheme with rateless codes for urllc," in *2020 IEEE 92nd Vehicular Technology Conference (VTC2020-Fall)*. IEEE, 2020, pp. 1–5.
- [24] W. Zhang, X. Zhong, *et al.*, "On analysis of asynchronous grant-free access with rateless codes," in *2022 IEEE Global Communications Conference (GLOBECOM)*, *in press*. IEEE.
- [25] B. Wang, L. Dai, *et al.*, "Dynamic compressive sensing-based multi-user detection for uplink grant-free noma," *IEEE Communications Letters*, 2016, vol. 20, no. 11, pp. 2320–2323.
- [26] Y. Du, B. Dong, *et al.*, "Efficient multi-user detection for uplink grant-free noma: Prior-information aided adaptive compressive sensing perspective," *IEEE Journal on Selected Areas in Communications*, 2017, vol. 35, no. 12, pp. 2812–2828.
- [27] D. L. Donoho, A. Maleki, *et al.*, "Message-passing algorithms for compressed sensing," *Proceedings of the National Academy of Sciences*, 2009, vol. 106, no. 45, pp. 18 914–18 919.
- [28] J. Ma and L. Ping, "Orthogonal amp," *IEEE Access*, 2017, vol. 5, pp. 2020–2033.
- [29] T. P. Minka, "Expectation propagation for approximate bayesian inference," *arXiv preprint arXiv:1301.2294*, 2013.
- [30] C. Wei, H. Liu, *et al.*, "Approximate message passing-based joint user activity and data detection for noma," *IEEE Communications Letters*, 2016, vol. 21, no. 3, pp. 640–643.
- [31] Y. Mei, Z. Gao, *et al.*, "Compressive sensing-based joint activity and data detection for grant-free massive iot access," *IEEE Transactions on Wireless Communications*, 2021, vol. 21, no. 3, pp. 1851–1869.
- [32] W. Zhu, M. Tao, *et al.*, "Deep-learned approximate message passing for asynchronous massive connectivity," *IEEE Transactions on Wireless Communications*, 2021, vol. 20, no. 8, pp. 5434–5448.
- [33] Z. Zhang, Y. Chi, *et al.*, "Asynchronous grant-free random access: Receiver design with partially uni-directional message passing and interference suppression analysis," *arXiv preprint arXiv:2305.09954*, 2023.
- [34] T. Ding, X. Yuan, *et al.*, "Sparsity learning-based multiuser detection in grant-free massive-device multiple access," *IEEE Transactions on Wireless Communications*, 2019, vol. 18, no. 7, pp. 3569–3582.
- [35] S. Jiang, C. Xu, *et al.*, "Bayesian receiver design for asynchronous massive connectivity," *IEEE Transactions on Wireless Communications*, 2022.
- [36] W. Zhang, J. Li, *et al.*, "A joint user activity detection and channel estimation scheme for packet-asynchronous grant-free access," *IEEE Wireless Communications Letters*, 2021, vol. 11, no. 2, pp. 338–342.
- [37] W. J. Lim, R. Abbas, *et al.*, "Analysis and design of analog fountain codes for short packet communications," *IEEE Transactions on Vehicular Technology*, 2021, vol. 70, no. 12, pp. 12 662–12 674.
- [38] X. Meng, S. Wu, *et al.*, "Approximate message passing with nearest neighbor sparsity pattern learning," *arXiv preprint arXiv:1601.00543*, 2016.
- [39] M. Ke, Z. Gao, *et al.*, "Compressive sensing-based adaptive active user detection and channel estimation: Massive access meets massive mimo," *IEEE Transactions on Signal Processing*, 2020, vol. 68, pp. 764–779.
- [40] Z. Zhang, X. Wang, *et al.*, "Grant-free rateless multiple access: A novel massive access scheme for internet of things," *IEEE Communications Letters*, 2016, vol. 20, no. 10, pp. 2019–2022.

Biographies



Wei Zhang received a B.S. degree in communication engineering from Beijing Jiaotong University in 2018. He is currently pursuing a Ph.D. degree with the Department of Electronic Engineering, Tsinghua University. His research interests are in the areas of random access, channel coding, and signal processing.



Xiaofeng Zhong received his Ph.D. degree in Information and Communication Systems from Tsinghua University in 2005. He has been an associate professor in the Department of Electronic Engineering at Tsinghua University. He researches in the field of mobile networks, including the users' behaviors and traffic model analysis, MAC and Network protocol design, and resource management optimization. Prof. Zhong has published more than 100 papers and owns 11 patents.



Shidong Zhou received B.S. and M.S. degrees from Southeast University, Nanjing, China, in 1991 and 1994, respectively, and the Ph.D. degree at Tsinghua University in 1998. He is currently a Professor at the Department of Electronic Engineering, at Tsinghua University. He was involved in the Program for New Century Excellent Talents in University 2005 (Ministry of Education). He received the Special Award of the 2016 National Prize for Progress in Science and Technology. His research interest lies in wireless transmission techniques, including distributed wireless communication systems, multiple access techniques, channel sounding and modeling, coordination of communication, control, and computing, and application in future mobile communications.