

RESEARCH

Open Access



Distributed ranking-based resource allocation for sporadic M2M communication

Yunyan Chang^{1*} , Peter Jung^{2,3}, Chan Zhou¹ and Sławomir Stańczak^{2,3}

*Correspondence:
yunyanchang@outlook.com

¹ Huawei European Research
Center, Riesstr. 25 C,
80992 Munich, Germany

² Fraunhofer Heinrich-Hertz-
Institute, Einsteinufer 37,
10587 Berlin, Germany

³ Technische Universität
Berlin, Straße des 17. Juni 135,
10623 Berlin, Germany

Abstract

This work proposes a novel scheme for distributed ranking-based and contention-free resource allocation in large-scale *machine-to-machine* (M2M) communication networks. We partition a network of N devices into disjoint clusters based on service type, and assign to each cluster a cluster-specific signature for active cluster members to indicate their active status. The devices in each cluster are totally ordered in some a priori-known manner, which gives rise to an *active ranking* of active cluster members. In order to tackle complexity issues in large-scale M2M networks with a massive number of devices, we propose a distributed resource allocation scheme using the framework of *compressed sensing* (CS), which mainly consists of three phases: (i) In a full-duplex acquisition phase, the devices transmit their cluster-specific signatures simultaneously and the network activation pattern is collected in a distributed manner. (ii) The base station detects the active clusters and the number of active devices per cluster using *block sketching*, and allocates resources to each active cluster accordingly. (iii) Each active device determines its active ranking in the cluster and accesses a specific resource according to the ranking position. By exploiting the sparsity in the activation pattern of the M2M devices, the proposed scheme is formulated as a CS support recovery problem for a particular binary block-sparse signal $x \in \mathbb{B}^N$ – with block sparsity K_B and in-block sparsity K_I over block size d . Our analysis shows that the proposed scheme efficiently reduces the signature length to $\mathcal{O}(\max\{K_B \log N, K_B K_I \log d\})$ and achieves less computational complexity of $\mathcal{O}(dK_I^2 + \frac{N}{d} \log N)$ compared with standard CS algorithms. Moreover, numerical results suggest strong robustness of the proposed scheme under noisy conditions.

Keywords: Compressed sensing, Block sparsity, M2M, Distributed resource allocation, Sketching

1 Introduction

Towards the next generation of mobile and wireless networks, *machine-to machine* (M2M) communications [3] is envisioned to play a significant role that forms the basis for the future *Internet of Things* (IoT). It is anticipated that the number of M2M devices connected to a single cellular *base station* (BS) may exceed 5 million by the year 2030 [4]. In such a massive connectivity scenario, conventional random access and resource allocation schemes [5] are not sufficient to handle the massive amount of communication

requests from the M2M devices, thereby posing significant challenges in the design and operation of the wireless networks.

For large-scale networks such as M2M communication networks, *distributed* resource allocation schemes are highly desired in order to alleviate the scaling issues due to massive connectivity [6]. In distributed schemes, each device decides semi-autonomously on which resources to access the channel, which results in a significant reduction of signaling overhead for coordination and information exchange between the devices and the BS. Moreover, considering the limited computing capability of the M2M devices [7], there is a strong need for low-complexity solutions that require relatively low computing power.

Apart from the massive number of devices, a key feature of the M2M traffic is that the device activity patterns are typically sporadic [8] so that at any given time instant each device has a low probability of being active. This results in a certain level of sparsity in the device activity. To this end, *compressed sensing* (CS) [9–12] is identified as an ideal framework in such scenarios since it provides tools for efficient reconstruction of high-dimensional signals with a sparse representation. Moreover, it is often observed that the messages sent by the M2M devices are strongly correlated, e.g., due to proximity, the same service type, and etc [7]. Therefore, it is in particular beneficial to partition a given set of all devices into a number of clusters such that similar requests can be handled jointly. Given this cluster-like behavior and the sparse nature of the M2M traffic, the activation pattern of the devices can be modeled as a block-sparse signal with an additional in-block structure [13] in CS-based applications.

The main objective of this study is to seek for an efficient resource allocation scheme to mitigate serious scaling problems resulting from massive connectivity, by exploiting the specific sparsity feature in the device activity. A special attention is attached to low overhead communication with enhanced spectral efficiency and reduced computational complexity. Towards this end, we propose a three-phase resource allocation scheme, where we use a distributed approach to reduce the communication overhead and a sketching algorithm to lower the computing load.

1.1 Related works

Conventional cellular networks are designed based on the scheduling of active users to orthogonal time or frequency resources. The excessive control overhead incurred by the massive number of sporadically active M2M devices, however, renders such kind of resource allocation schemes unrealistic. As an alternative, contention-based schemes, such as slotted ALOHA [14], have been proposed to deal with this issue. However, the main drawback of these schemes is the lack of deterministic performance guarantees, and in particular the performance deteriorates significantly under massive connectivity settings. In addition, the authors of [15] also raised concerns of energy efficiency in large-scale M2M networks, where systems with less computational complexity and lower power consumption are particularly desired for low-cost M2M devices with limited operational capability.

Taking sparsity in the activity pattern into account, access schemes using CS also receive a great deal of attention in recent years. To the best of our knowledge, the authors of [16] were the first to propose the idea of CS-based multi-user detection. They

introduced a smart adaptive algorithm that switches between a CS-based reconstruction algorithm and a classical detection method depending on the sparsity level of the signals being detected. Reference [17] proposed greedy CS algorithms to facilitate a joint detection of device activity and transmitted data. The idea was further extended in [18] to include multicarrier access schemes and to provide higher spectral efficiency and more flexibility of such schemes. Furthermore, schemes for distributed compressed sensing were also widely studied (e.g. in [19]) to take advantage of both inter- and intra-signal correlations by jointly reconstructing signals that have been compressed independently. The concept of distributed compressed sensing has been applied by the authors of [20] to facilitate device detection in M2M communications, which shows significant performance gains expressed in terms of robustness.

However, none of the works mentioned above has exploited the particular cluster-like behavior of M2M devices, thereby ignoring the block sparsity structure in the activation pattern. In addition, greedy algorithms such as *orthogonal matching pursuit* (OMP) [21] are widely used in detection schemes owing to their low computational complexity. However, in many wireless applications including that considered in this study, even this reduced complexity is a bottleneck due to strictly limited computing resources on the M2M devices. Therefore, we employ a sketching algorithm in our proposed approach to further offload the computation burden incurred by massive M2M communications.

1.2 Main contribution

In this work, we present and analyze a novel distributed scheme for device detection and contention-free resource allocation in large-scale M2M communication networks. We partition the M2M devices into disjoint clusters in advance based on service type, and assign to each device a cluster-specific signature that active devices use for their initial access. In this paper, we use the following definition of an active cluster.

Definition 1 (Active cluster) We say that a cluster is active if there is at least one active device in this cluster.

The devices in each cluster are *totally ordered* according to some given criterion such as service priority. In other words, the set of devices in each cluster is a totally ordered set \mathcal{S} so that if $a, b \in \mathcal{S}, a \neq b$ and $a \leq b$, then $a < b$. This order is a priori known at all cluster members and gives rise to what we call *active ranking* that determines the order within each cluster in which the active devices access the set of assigned resource blocks, i.e., an active device that is the i -th element in the active ranking accesses the i -th resource block.

Definition 2 (Active ranking) The active ranking associated with each active cluster is the totally ordered subset induced by the active devices in a cluster.

Motivated by the CS principle, the proposed scheme mainly consists of three phases:

- *Phase (i) Signal acquisition* The active devices transmit simultaneously pre-equalized individual signatures, each of which indicates the membership to a particular cluster.

Exploiting full-duplex transceivers,¹ all the devices and the BS obtain their individual measurements, which are the superposition of the transmitted signatures from the active devices.

- *Phase (ii) Detection at BS* The BS detects the active clusters, the number of active devices in each cluster, and also the collision pattern in the received measurements. Then it broadcasts the detected information to the devices and assigns a sufficient amount of resources to each active cluster accordingly.
- *Phase (iii) Detection at devices* Each active device detects the active ranking of its cluster, and then accesses the corresponding resource assigned by the BS based on its ranking position.

We study a particular signature design for the devices in each cluster to facilitate the detection process of phase (ii) at the BS side and that of phase (iii) at the device side. Moreover, based on the Count-Sketch procedure [23, 24], we develop a novel block sketching algorithm to perform phase (ii) and to bring down computational complexity induced by massive connectivity. Phase (iii) is performed using a conventional greedy algorithm such as OMP [21] except that we use feedback information from the BS to enhance robustness and to further reduce complexity. Furthermore, with the distributed ranking-based resource allocation scheme, each active device accesses autonomously a pool of resource blocks assigned by the BS in a contention-free manner without further control signaling, thus the communication overhead is significantly reduced. We show via theoretical analysis that the proposed scheme achieves a better scaling with increasing network size both in terms of communication overhead and computational complexity when compared with classical CS-based approaches. Moreover, the simulation results reveal a significantly enhanced robustness of the proposed scheme in the presence of Gaussian noise and inaccurate channel estimations.

1.3 Organization of the paper

The remainder of this paper is organized as follows. Section 2 introduces a mathematical model for the proposed scheme. In Sect. 3, we present the detection algorithms in detail. Section 4 is devoted to theoretical analysis of the proposed scheme, while Sect. 5 evaluates the performance with numerical simulation results. Finally, Sect. 6 concludes the paper with some final remarks.

1.4 Notational remarks

Throughout this work, matrices and vectors are denoted by uppercase and lowercase letters, respectively. The superscript $(\cdot)^T$ represents the transpose of a matrix or a vector, and $(\cdot)^H$ indicates the Hermitian transpose. $A \circ B$ denotes the Hadamard product of matrices A and B . The field of binary, real and complex numbers are denoted by \mathbb{B} , \mathbb{R} and \mathbb{C} . The cardinality of a set is given by $|\cdot|$, and the ℓ_2 -norm is given by $\|\cdot\|$. Furthermore, \mathcal{O} denotes “big-O” according to Knuth’s notation. Unless otherwise stated, all logarithms are assumed to be to base 2.

¹ The authors in [22] highlighted benefits of the full duplex wireless and rendered its feasibility for implementation in future IoT systems.

2 Methods

In this section, we introduce the underlying system model and formulate the problem which is addressed in this paper.

2.1 Transmitter side

We consider an M2M network consisting of N devices, which are partitioned into L clusters of equal size d according to the service type.² The members of each cluster are known both at the BS and at all devices, which can be achieved via device registration to the network. We assume that the devices have sporadic transmissions, which implies that at a given time instant, only a relatively small number of devices from a few clusters are activated to access the channel. Therefore, we define a twofold sparsity pattern to model the active status of the devices:

- *Block sparsity* K_B The maximum number of active clusters at any time.
- *In-block sparsity* K_I The maximum number of active devices in an active cluster.

Therefore, the total number of active devices in the network is $K \leq K_B K_I$. On the other hand, due to the sparsely-activated devices, we have $K \leq K_B K_I \ll N = Ld$.

We use a K -sparse binary sequence (or vector) $x \in \mathbb{B}^N$ to model the activation pattern of the devices in the network, where entry “1” indicates that the corresponding device is active while an inactive device results in “0”. Furthermore, we use $x_\ell \in \mathbb{B}^d, \ell \in \{1, \dots, L\}$ to denote the subsequence (or subvector) of x corresponding to cluster ℓ . In addition, the *block support*, denoted as \mathcal{S}_B , is defined to be the set of indices of the active clusters: $\mathcal{S}_B = \{\ell \in \{1, \dots, L\} : \|x_\ell\|_0 \neq 0\}$. Similarly, the *in-block support*, denoted as $\mathcal{S}_{I,\ell}$, indicates the set of indices of the active devices in cluster ℓ : $\mathcal{S}_{I,\ell} = \{j \in \{1, \dots, d\} : x_{\ell,j} = 1\}$. By definition, we have $|\mathcal{S}_B| \leq K_B$ and $|\mathcal{S}_{I,\ell}| \leq K_I$ for all $\ell \in \{1, \dots, L\}$. Thus, the activation pattern of the devices x is modeled as a sparse signal with block sparsity K_B and in-block sparsity K_I , and we call signals of such a sparsity pattern (K_B, K_I) *block-sparse*.

Due to the sparsity in x , we use the CS theory [9–11] to reconstruct x based on measurements performed by the BS and the devices in the network. We use $A \in \mathbb{R}^{M \times N}$ to denote the measurement matrix whose exact structure is defined later in Sect. 3.1. Each column of A , say column i denoted by $a_i, i \in \{1, \dots, N\}$, corresponds to the unique signature sent by device i if it is active, whereas $A_{-, \ell} \in \mathbb{R}^{M \times d}$ denotes a submatrix of A corresponding to the signatures sent by the devices from the ℓ -th cluster.

2.2 Receiver side

Due to the superposition property of the wireless channel, each receiver observes a noisy superposition of signatures transmitted by the active devices. We assume that all frames are received synchronously at the aggregation node. In practical systems this can be ensured e.g. by a synchronization signal from the BS. Alternatively, the requirement for precise synchronization can be mitigated by using schemes such as that considered in [25] which is robust to the lack of synchronization and requires only a coarse synchronization. To this end, the received signal at the BS $y \in \mathbb{C}^M$ is a noisy linear combination of the transmitted signatures given by

² The scheme can be easily extended to unequal-sized clusters. We only use clusters of equal size for simplicity.

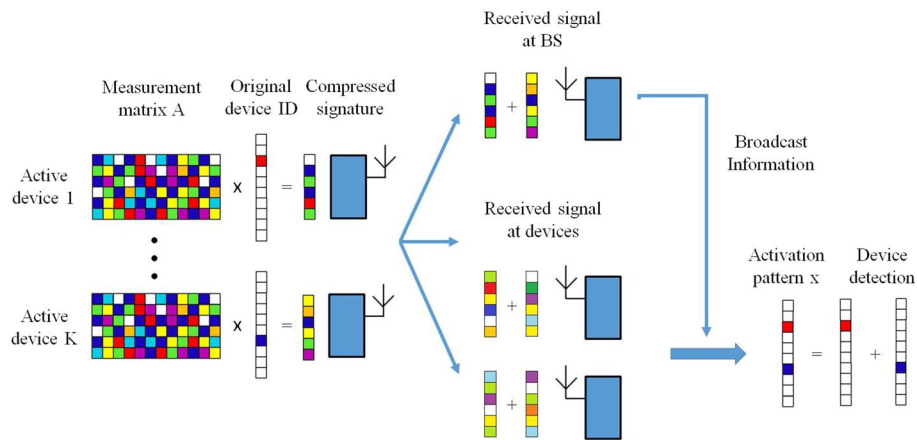


Fig. 1 Target problem formulation

$$y = (A \circ H_B)x + \epsilon_B, \quad (1)$$

where $H_B \in \mathbb{C}^{M \times N}$ is the channel matrix, and $\epsilon_B \in \mathbb{C}^M$ denotes the additive noise which is assumed to be zero-mean with independent components of variance σ_B^2 . The Hadamard product is used here to model the effective channel when applied with advanced technologies such as frequency hopping [26], which is introduced in detail in Sect. 3.2.

In addition, the active devices also perform their own local measurements during the acquisition phase. The received signal $y_D \in \mathbb{C}^M$ observed by a particular device at some given time instant is obtained as

$$y_D = (A \circ H_D)x + \epsilon_D, \quad (2)$$

where H_D is an $M \times N$ matrix representing the wireless channels between the devices, and $\epsilon_D \in \mathbb{C}^M$ is the vector of independent noise components with zero mean and variance σ_D^2 .

Furthermore, we make the following assumption on channel knowledge:

Assumption 1 Each device has the channel state information to the BS.

In current systems [5], the channel information can be obtained with pilot-signals from the network with no extra cost. Alternatively, it can also be possibly obtained via statistical channel knowledge [27], location-based estimation [28], channel reciprocity [29] or long-term observation [30]. We point out that the entries of H_B and H_D are in general not physical channels but rather effective channels that depend on the underlying transmission scheme. For instance, the effective channels of the energy-detection based scheme in [25] is related to channel power gains, which are much easier to acquire than complex-valued channel coefficients.

2.3 Problem formulation

Figure 1 illustrates the transmission scenario and the problem of interest in this work. As a result of concurrent transmissions of the active devices, the BS observes a noisy superposition of the transmitted signatures given by (1), while the received signal at the

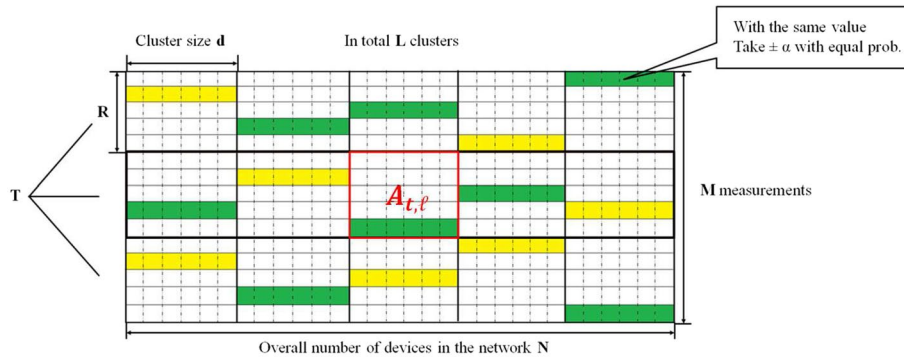


Fig. 2 Structure of the measurement matrix A

devices is as in (2). Hereafter, we apply CS related techniques to reconstruct the (K_B, K_I) block-sparse signal x with y at the BS and y_D at the devices. To be specific, the BS performs *block support recovery* to obtain an accurate estimation of the block support \tilde{S}_B and the cardinality of the in-block support $|\tilde{S}_{I,\ell}|$ for each $\ell \in \{1, \dots, L\}$. Subsequently, each active device from cluster ℓ performs *in-block support recovery* to estimate the in-block support $\tilde{S}_{I,\ell}$ using the received signal y_D in (2) and side information broadcast by the BS.

In a typical massive connectivity scenario as in M2M communications, we need to find proper solutions of high efficiency and low computational complexity to the following problems.

- *Problem 1* Block support recovery at the BS with $\mathbb{P}(\tilde{S}_B = S_B, \forall \ell \in \{1, \dots, L\} |\tilde{S}_{I,\ell}| = |S_{I,\ell}|) \geq 1 - \delta$, where δ is a target error bound.
- *Problem 2* In-block support recovery at the active devices in cluster ℓ with $\mathbb{P}(\tilde{S}_{I,\ell} = S_{I,\ell}) \geq 1 - \delta$.³

3 Distributed ranking-based resource allocation

In this section, we present our algorithm design to tackle the target problems. In particular, this includes the structured signature model and the decoding procedures at the BS as well as at the devices, respectively.

3.1 Structured random signature model

The measurement matrix $A \in \mathbb{R}^{M \times N}$ we design here is a structured random matrix, which is an extension of those utilized by the Count-Sketch procedure proposed in [23, 24]. These matrices are desired as they in general facilitate low computational complexity. We denote by $\mathcal{A}(R, T, L, d, \alpha)$ a particular distribution over matrices with RT rows and Ld columns, which is specified below, and we assume that the measurement matrix A is drawn from this distribution, i.e., $A \sim \mathcal{A}(R, T, L, d, \alpha)$ with $M = RT, N = Ld$.

As illustrated in Fig. 2, the measurement matrix A is composed of the vertical concatenation of T individual random matrices that we denote by $A_{t,-} \in \mathbb{R}^{R \times N}$ for

³ For simplicity we use the same target error bound δ for Problem 1 and 2.

$t \in \{1, \dots, T\}$, where $A_{t,-}$ consists of the horizontal concatenation of L sub-matrices $A_{t,\ell} \in \mathbb{R}^{R \times d}$ for $\ell \in \{1, \dots, L\}$. Each $A_{t,\ell}$ is a sparse matrix containing exactly d non-zero components – located on the same row and with the same value. The index of the row with non-zero elements is chosen uniformly at random from the set $\{1, 2, \dots, R\}$, and the non-zero components take either the value $+\alpha$ or $-\alpha$ with probability $1/2$. For a given realization of $A_{t,\ell}$, let $q_{t,\ell} \in \{1, \dots, R\}$ denote the index of the row of $A_{t,\ell}$ with non-zero entries, and $s_{t,\ell} \in \{-\alpha, +\alpha\}$ be the corresponding value of the non-zero components in $A_{t,\ell}$. As each signature transmitted by the devices is the corresponding column of A , it is therefore a sparse sequence of length M with sparsity level T .

3.2 Block support recovery at BS

The objective of the decoding procedure at the BS is to obtain an accurate estimation of the block support \mathcal{S}_B and the cardinality of the in-block support $|\mathcal{S}_{I,\ell}|$ for each active cluster $\ell \in \{1, \dots, L\}$ (see *Problem 1* in Sect. 2.3).

The signal y received by the BS at some given time instant is given by (1). Since the channel state information H_B is assumed to be available at the devices, the active devices can perform a channel inversion before transmitting the signatures to indicate their active state. Particular variants of (generalized) inverses of the channel matrix may be taken at the transmitter side. For example, H_B can be inverted by simply taking the reciprocal of the non-zero elements whose magnitude is above certain threshold θ , which is given by

$$(H_B)_{ij}^{-1} = \begin{cases} \frac{1}{(H_B)_{ij}}, & \text{if } |H_B|_{ij} \geq \theta, \\ 0, & \text{else} \end{cases}, \text{ for } i = \{1, \dots, M\}, j = \{1, \dots, N\}. \quad (3)$$

In this case devices with weak links due to the near-far behavior may stay offline and avoid excessively large transmit power and strong interference to other nodes. Then the obtained measurements at the BS are given as

$$y = (A \circ H_B \circ H_B^{-1})x + \epsilon_B := Ax + \epsilon_B. \quad (4)$$

In this paper, for the sake of simplicity we develop a fast block sketching algorithm based on the Count-Sketch procedure proposed in [23, 24] to realize the decoding process at the BS. To be more precise, we use $y_t \in \mathbb{R}^R$ to denote the subvector of y corresponding to the observations obtained via submatrix $A_{t,-}$. So, we have

$$y_t = A_{t,-}x + \epsilon_t, \quad \text{for } t \in \{1, \dots, T\}. \quad (5)$$

Given y_t for each $t \in \{1, \dots, T\}$, we form the signal estimate $\tilde{x}_t \in \mathbb{R}^N$ by indexing and scaling the entries of the corresponding observations y_t such that

$$\tilde{x}_t = A_{t,-}^T y_t, \quad \text{for } t \in \{1, \dots, T\}, \quad (6)$$

where each $A_{t,-}$ consists of the horizontal concatenation of L submatrices $A_{t,\ell}$ for $\ell \in \{1, \dots, L\}$. Further recall that each matrix $A_{t,\ell}$ is a sparse matrix containing d non-zero components located on the same row $q_{t,\ell} \in \{1, \dots, R\}$ and all the non-zero

components have the same value $s_{t,\ell} \in \{-\alpha, +\alpha\}$. As a result, the i -th entry $\tilde{x}_{t,i}$ of \tilde{x}_t for the ℓ -th block can be written as

$$\tilde{x}_{t,i} = s_{t,\ell} y_{t,\ell}, \quad \text{for } i \in \{d\ell - d + 1, \dots, d\ell\}, \ell \in \{1, \dots, L\}, \quad (7)$$

where $y_{t,\ell} = (y_t)_{q_{t,\ell}}$. We use $a_{q_{t,\ell},i}$ to denote the entry in the $q_{t,\ell}$ -th row and i -th column of $A_{t,-}$, then we have

$$y_{t,\ell} = \sum_{k=1}^L \sum_{i=d\ell-d+1}^{dk} a_{q_{t,\ell},i} x_i + \epsilon_{t,\ell}, \quad \text{for } t \in \{1, \dots, T\}, \ell \in \{1, \dots, L\}. \quad (8)$$

Since $y_{t,\ell}$ is dominated by the non-zero entries on the $q_{t,\ell}$ -th row of $A_{t,-}$, we denote by $\mathcal{S}_{t,\ell}$ a set of indices of the clusters which have non-zero components on the same row as block ℓ , i.e., $\mathcal{S}_{t,\ell} = \{j = \{1, \dots, L\} : q_{t,j} = q_{t,\ell}\}$. Then $y_{t,\ell}$ is given by

$$\begin{aligned} y_{t,\ell} &= \sum_{k \in \mathcal{S}_{t,\ell}} \sum_{i=d\ell-d+1}^{dk} s_{t,k} x_i + \epsilon_{t,\ell} \\ &= \sum_{i=d\ell-d+1}^{d\ell} s_{t,\ell} x_i + \sum_{k \in \mathcal{S}_{t,\ell} \setminus \{\ell\}} \sum_{i=d\ell-d+1}^{dk} s_{t,k} x_i + \epsilon_{t,\ell} \\ &\stackrel{(a)}{=} s_{t,\ell} \left(\sum_{i=d\ell-d+1}^{d\ell} x_i \right) + \sum_{k \in \mathcal{S}_{t,\ell} \setminus \{\ell\}} s_{t,k} \left(\sum_{i=d\ell-d+1}^{dk} x_i \right) + \epsilon_{t,\ell} \\ &\stackrel{(b)}{=} s_{t,\ell} |\mathcal{S}_{I,\ell}| + \sum_{k \in \mathcal{S}_{t,\ell} \setminus \{\ell\}} s_{t,k} |\mathcal{S}_{I,k}| + \epsilon_{t,\ell} \\ &\stackrel{(c)}{=} s_{t,\ell} |\mathcal{S}_{I,\ell}| + \sum_{k \in \mathcal{S}_B \cap \mathcal{S}_{t,\ell} \setminus \{\ell\}} s_{t,k} |\mathcal{S}_{I,k}| + \epsilon_{t,\ell}, \end{aligned} \quad (9)$$

where (a) follows from the structure of $A_{t,\ell}$ with equal non-zero elements on the same row, (b) is due to $x_i \in \{0, 1\}$, and (c) holds since $|\mathcal{S}_{I,k}| = 0$ if $k \notin \mathcal{S}_B$.

Putting $y_{t,\ell}$ in (9) into (7) yields

$$\begin{aligned} \tilde{x}_{t,i} &= s_{t,\ell} \left(s_{t,\ell} |\mathcal{S}_{I,\ell}| + \sum_{k \in \mathcal{S}_B \cap \mathcal{S}_{t,\ell} \setminus \{\ell\}} s_{t,k} |\mathcal{S}_{I,k}| + \epsilon_{t,\ell} \right) \\ &= \underbrace{\alpha^2 |\mathcal{S}_{I,\ell}| + \sum_{k \in \mathcal{S}_B \cap \mathcal{S}_{t,\ell} \setminus \{\ell\}} s_{t,\ell} s_{t,k} |\mathcal{S}_{I,k}|}_{\Delta_{t,\ell}} + \underbrace{s_{t,\ell} \epsilon_{t,\ell}}_{\Xi_{t,\ell}}, \end{aligned} \quad (10)$$

where $\Delta_{t,\ell}$ is the interference term from other active blocks, and $\Xi_{t,\ell}$ is the noise term with zero mean and variance γ^2 .

To mitigate the interference $\Delta_{t,\ell}$ from other blocks, we consider a block-wise estimate \bar{x}_ℓ for each $\ell \in \{1, 2, \dots, L\}$ given by

$$\bar{x}_\ell = \text{median}\{\tilde{x}_{t,i}\}_{t=1, i=d\ell-d+1}^{T, d\ell}, \quad \text{for } \ell \in \{1, 2, \dots, L\}. \quad (11)$$

Notice that, instead of the mean, the estimate \bar{x}_ℓ for block ℓ is equal to the median of $\tilde{x}_{t,i}$ over $\mathcal{O}(Td)$ samples. The rationale behind this approach is to make the estimates more

robust against outliers, since large value elements in the data stream may spoil some subsets of the estimate if the mean is computed.

We show later in Sect. 4.1 that by taking the median value block-wisely among all individual estimations as in (11), each estimate \tilde{x}_ℓ for the ℓ -th block corresponds to $|S_{I,\ell}|$ with high probability (w.h.p.), and the cardinality of the in-block support set $|S_{I,\ell}|$ is obtained as

$$|S_{I,\ell}| = \left\lceil \frac{1}{\alpha^2} \cdot \tilde{x}_\ell \right\rceil, \quad \text{for } \ell \in \{1, 2, \dots, L\}. \quad (12)$$

Therefore, since $|S_{I,\ell}|$ indicates the number of active devices in cluster $\ell \in \{1, 2, \dots, L\}$, those clusters with $|S_{I,\ell}| > 0$ are marked as “active” by the BS. For brevity, we assume that each device needs one resource block for transmission. Therefore, the number of resource blocks assigned by the BS to cluster $\ell \in \{1, 2, \dots, L\}$ equals to $|S_{I,\ell}|$.

In addition, for a given x_i from an active block $\ell \in \mathcal{S}_B$, if an individual estimate $\tilde{x}_{t,i}$ in (7) is much larger than the block-wise estimate \tilde{x}_ℓ in (11), i.e., $\tilde{x}_{t,i} \gg \tilde{x}_\ell$, then we can conclude that the corresponding measurement might suffer strong interference from the other active clusters. That is, for a given x_i from block ℓ and a particular $t \in \{1, \dots, T\}$, the interference term $\Delta_{t,\ell}$ in (10) is non-zero. In such a case, we mark the measurement as “collided” for cluster ℓ and keep its index $q_{t,\ell}$ in the collision pattern vector Q_ℓ for the corresponding cluster.

The above approach provides the BS with an accurate estimate of \mathcal{S}_B and $|S_{I,\ell}|$, and therefore it solves *Problem 1*. The detailed proof of the performance guarantee will be given in Sect. 4.1. In addition, it provides the collision patterns Q_ℓ in the measurements for $\ell \in \{1, 2, \dots, L\}$. The BS broadcasts these information to the devices and assigns to each cluster, say cluster $\ell \in \{1, 2, \dots, L\}$, $|S_{I,\ell}|$ resource blocks to accommodate all active devices in this cluster.

3.3 In-block support recovery at devices

During the signal acquisition phase, each active device also collects its own measurements, which are linear combinations of the transmitted signatures from other active devices. In this section, we address *Problem 2* in Sect. 2.3. The objective is to develop a scheme that enables each active device in cluster $\ell \in \{1, \dots, L\}$ to reliably estimate the in-block support $S_{I,\ell}$ with low computational complexity, based on its local measurements and the feedback from the BS as side information.

Given the measurement matrix A under the random structured model in Sect. 3.1 and the pre-channel correction in (4), the measurement y_D collected at an active device is given by

$$y_D = (A \circ H_D \circ H_B^{-1})x + \epsilon_D = \tilde{A}x + \epsilon_D. \quad (13)$$

According to the specific structure of A , the corresponding submatrix $A_{-, \ell}$ for cluster $\ell \in \{1, \dots, L\}$ has only T rows with non-zero components. The indices of these rows are collected in the set D_ℓ . Furthermore, with the feedback information from the BS on the collision pattern Q_ℓ for cluster ℓ indicating those collided measurements to be discarded, we form an index set $U_\ell = D_\ell \cap \bar{Q}_\ell$. Therefore, in order to perform the in-block support

recovery of x_ℓ at any device from cluster ℓ , we simply need to focus on $y_{D,\ell}$ – a vector composed of the entries of y_D corresponding to U_ℓ . We denote $\tilde{A}_{D,\ell}$ as a $|U_\ell| \times d$ submatrix of \tilde{A} with vertical concatenation of rows corresponding to U_ℓ and columns for block ℓ . Therefore, we have

$$y_{D,\ell} = \tilde{A}_{D,\ell} x_\ell + \tilde{\epsilon}_D. \quad (14)$$

As introduced in Sect. 2.2, we use technologies such as frequency hopping [26] for the transmission where symbols are transmitted hopping over multiple subcarriers. Since the channels between the devices over different subcarriers are assumed to be i.i.d., we can conclude that $\tilde{A}_{D,\ell}$ has independent columns and row-blocks (e.g., to be i.i.d Subgaussian). Therefore, some classic CS decoding algorithms can be applied to perform the in-block support recovery. We argue in favor of the greedy algorithms such as OMP [21] due to their low complexity, which is particularly attractive to M2M-based applications where limited computational capability as well as energy consumption at the devices are important design criteria.

An example of the modification on OMP for in-block support recovery is summarized in Algorithm 1, where the modified steps are marked in boldface. By exploiting the broadcast information on the cardinality of the in-block support $|S_{I,\ell}|$ for cluster ℓ , the stopping criteria for implementing the greedy algorithms can be set by limiting the number of iterations to $|S_{I,\ell}|$, thereby leading to further reduced computational complexity. This is merely an illustrative example assuming that channel knowledge is available at the devices. However, this assumption can be further relaxed by using *approximate message passing* (AMP) algorithms as in [31].

Algorithm 1 Modified OMP for In-block support recovery

Input: \tilde{A} , y_D , D_ℓ , Q_ℓ , and $|S_{I,\ell}|$.

Output: $S_{I,\ell}$.

- 1: **Form an index set** $U_\ell = D_\ell \cap \bar{Q}_\ell$.
- 2: **Construct a subvector** $y_{D,\ell}$ **of** y_D **composed of the entries corresponding to** U_ℓ , **and a submatrix** $\tilde{A}_{D,\ell}$ **of** \tilde{A} **with vertical concatenation of rows corresponding to** U_ℓ **and columns corresponding to block** ℓ .
- 3: Initialize the residual $r_0 = y_{D,\ell}$, the index set $\Lambda_0 = \emptyset$, the matrix of the chosen atoms $\Phi_0 = \emptyset$, and the iteration counter $k = 1$.
- 4: Choose the column of $\tilde{A}_{D,\ell}$ with index λ_k that is best matched to r_{k-1} according to

$$\lambda_k = \arg \max_{a_\lambda \in \tilde{A}_{D,\ell}} |\langle r_{k-1}, a_\lambda \rangle|. \quad (15)$$

- 5: Augment $\Lambda_k = \Lambda_{k-1} \cup \{\lambda_k\}$ and the matrix of the chosen atoms $\Phi_k = [\Phi_{k-1} \ a_{\lambda_k}]$.
- 6: Solve the least square error minimization problem to obtain a new signal estimate:

$$x_{\ell,k} = \arg \min_{x_\ell} \|y_{D,\ell} - \Phi_k x_\ell\|_2. \quad (16)$$

- 7: Update the residual as $r_k = y_{D,\ell} - \Phi_k x_{\ell,k}$.
 - 8: **Increment** k **by 1, and return to Step 4 until** $k > |S_{I,\ell}|$.
 - 9: **return** $S_{I,\ell} = \Lambda_k$.
-

To this end, *Problem 2* is explicitly resolved. In Sect. 4.2, we prove conditions under which the in-block support $S_{I,\ell}$ can be accurately reconstructed at the devices in cluster ℓ . Thus, the activation pattern x_ℓ of the ℓ -th cluster can be precisely reconstructed and detected by the active devices. Thereafter, each active device is able to learn the active ranking in its cluster and accesses the corresponding resource blocks assigned by the BS.

4 Theoretical performance analysis

This section provides a sufficient condition for the performance guarantees of our proposed scheme. In particular, we come with the following theorem.

Theorem 1 Suppose that the activation pattern of the devices is modeled as a (K_B, K_I) block-sparse signal $x \in \mathbb{B}^N$ over block size d , and the signatures transmitted by the devices are modeled as a measurement matrix $A \in \mathbb{R}^{M \times N}$ following the structure designed in Sect. 3.1. By applying the block sketching algorithm in Sect. 3.2 for decoding at the BS and the modified OMP in Sect. 3.3 for decoding at the devices, x can be reliably reconstructed by the proposed scheme with computational complexity of $\mathcal{O}(dK_I^2 + \frac{N}{d} \log N)$ if the length of the transmitted signatures

$$M = \mathcal{O}(\max\{K_B \log N, K_B K_I \log d\}). \quad (17)$$

A rigorous proof of Theorem 1 will be presented in the following, considering both the block support recovery at the BS and in-block support recovery at the devices, respectively. And we also show that it achieves a better scaling when compared with classical CS-based approaches both in terms of communication overhead and computational complexity.

4.1 Block support recovery at BS

First, we analyze the recovery guarantee for the individual estimate $\tilde{x}_{t,i}$ in (7).

Lemma 1 Suppose that $x \in \mathbb{B}^N$ is a (K_B, K_I) block-sparse signal over block size d , and $A \in \mathbb{R}^{M \times N}$ is randomly drawn from $\mathcal{A}(R, T, L, d, \alpha)$. Given $y \in \mathbb{R}^M$ in (4) and the estimate $\tilde{x}_{t,i}$ in (7) for a particular entry x_i from block $\ell \in \{1, \dots, L\}$ and a given $t \in \{1, \dots, T\}$, let $\Gamma(\tilde{x}_{t,i}) := \{|\tilde{x}_{t,i} - \alpha^2 |S_{I,\ell}| \leq 3\gamma\}$, where γ^2 is the variance of the noise term $\Xi_{t,\ell}$ in (10). The probability of $\Gamma(\tilde{x}_{t,i})$ is lower bounded by

$$\mathbb{P}(\Gamma(\tilde{x}_{t,i})) \geq 1 - \frac{K_B - 1}{R}. \quad (18)$$

1 Proof

According to (10), for a particular estimate $\tilde{x}_{t,i}$ of x_i in block ℓ with $t \in \{1, \dots, T\}$, $\Gamma(\tilde{x}_{t,i})$ holds w.h.p. if the corresponding interference term $\Delta_{t,\ell} = 0$, since the noise term $\epsilon_{t,\ell}$ is randomly drawn from a Gaussian ensemble with zero mean and variance γ^2 [32]. A sufficient (but not necessary) condition for $\Delta_{t,\ell} = 0$ to hold is that the set $\mathcal{S}_B \cap \mathcal{S}_{t,\ell} \setminus \{\ell\} = \emptyset$, where $\mathcal{S}_{t,\ell} = \{j = \{1, \dots, L\} : q_{t,j} = q_{t,\ell}\}$. This implies that $q_{t,\ell}$ is distinct from $q_{t,\bar{\ell}}$ for all $\bar{\ell} \in \mathcal{S}_B \setminus \{\ell\}$. Therefore, we have

$$\begin{aligned} \mathbb{P}(\Gamma(\tilde{x}_{t,i})) &\geq \mathbb{P}(\Delta_{t,\ell} = 0) \geq \mathbb{P}(\mathcal{S}_B \cap \mathcal{S}_{t,\ell} \setminus \{\ell\} = \emptyset) \\ &= \mathbb{P}\left(\forall_{\bar{\ell} \in \mathcal{S}_B \setminus \{\ell\}} : q_{t,\ell} \neq q_{t,\bar{\ell}}\right) \stackrel{(a)}{=} [\mathbb{P}(q_{t,\ell} \neq q_{t,\bar{\ell}})]^{K_B-1} \\ &= [1 - \mathbb{P}(q_{t,\ell} = q_{t,\bar{\ell}})]^{K_B-1} = \left(1 - \frac{1}{R}\right)^{K_B-1} \\ &\stackrel{(b)}{\geq} 1 - \frac{1}{R}(K_B - 1), \end{aligned} \quad (19)$$

where (a) follows since the index of rows with non-zero entries $q_{t,\ell}$ are drawn i.i.d. uniformly at random for each $\ell \in \{1, \dots, L\}$ and $|\mathcal{S}_B| = K_B$, and the inequality in (b) follows from the Bernoulli's inequality [33]. \square

Then we look into the recovery guarantee for the block estimate \bar{x}_ℓ in (11) obtained via the median operator.

Lemma 2 Suppose that $x \in \mathbb{B}^N$ is a (K_B, K_I) block-sparse signal over block size d , and $A \in \mathbb{R}^{M \times N}$ is randomly drawn from $\mathcal{A}(R, T, L, d, \alpha)$. Given $y \in \mathbb{R}^M$ in (4) and the block estimate \bar{x}_ℓ in (11) for a particular block $\ell \in \{1, \dots, L\}$, let $\Gamma(\bar{x}_\ell) := \{|\bar{x}_\ell - \alpha^2 \mathcal{S}_{I,\ell}| \leq 3\gamma\}$, where γ^2 is the variance of the noise term $\Xi_{t,\ell}$ in (10). The probability of $\Gamma(\bar{x}_\ell)$ is lower bounded by

$$\mathbb{P}(\Gamma(\bar{x}_\ell)) \geq 1 - \frac{\delta}{L}, \quad (20)$$

if $R = \mathcal{O}(K_B)$ and $T = \mathcal{O}\left(\log \frac{N}{\delta}\right)$, where $0 < \delta < 1$ is an arbitrary target error bound.

1 Proof

As in (11), the block estimate \bar{x}_ℓ is obtained by taking the median of the individual estimate $\tilde{x}_{t,i}$ over $\mathcal{O}(Td)$ samples for a given $\ell \in \{1, \dots, L\}$. Suppose at least $\frac{Td}{2}$ estimates $\tilde{x}_{t,i}$ fulfills the Γ condition of Lemma 1, then $\Gamma(\bar{x}_\ell)$ in Lemma 2 will follow affirmatively. We analyze in the following where $\Gamma(\tilde{x}_{t,i})$ holds for at least $\frac{Td}{2}$ individual estimates.

Let X_1, \dots, X_{Td} be independent (0,1) Bernoulli random variables where $X_t, 1 \leq t \leq Td$ indicates whether the corresponding estimate $\tilde{x}_{t,i}$ of x_i satisfies the Γ condition of Lemma 1. As proved in Lemma 1, the probability of each X_t being equal to 1 is $p \geq 1 - \frac{K_B-1}{R}$. Then the probability that the number of simultaneous occurrence of the events $\{X_t = 1\}$ exceeds $Td/2$ is given by [32]

$$\mathbb{P}\left(\sum_{t=1}^{Td} X_t > \frac{Td}{2}\right) = \sum_{t=\lfloor \frac{Td}{2} \rfloor + 1}^{Td} \binom{Td}{t} p^t (1-p)^{Td-t}. \quad (21)$$

A lower bound on this probability can be calculated using Chernoff's inequality [34] to obtain

$$\mathbb{P}\left(\sum_{t=1}^{Td} X_t > \frac{Td}{2}\right) = \mathbb{P}\left(\frac{1}{Td} \sum_{t=1}^{Td} X_t > \frac{1}{2}\right) \geq 1 - e^{-\frac{Td}{2p}(p-\frac{1}{2})^2}. \quad (22)$$

The minimum bound can be easily found as achieved by $p = 1/2$. By setting a lower threshold $\theta \in (\frac{1}{2}, 1)$ to the probability p , we have $p \geq 1 - \frac{K_B-1}{R} \geq \theta$ from which it follows that

$$R \geq \frac{1}{1-\theta}(K_B - 1) = \mathcal{O}(K_B). \quad (23)$$

⁴ Furthermore, (22) also implies that the lower bound of the probability scales as $1 - e^{-\mathcal{O}(Td)}$. By taking $T = \log\left(\frac{Ld}{\delta}\right) = \mathcal{O}\left(\log\frac{N}{\delta}\right)$ into (22) proves Lemma 2. \square

Given the proof of Lemma 2, the overall performance guarantee for the block support recovery at the BS follows inherently.

Lemma 3 Suppose that $x \in \mathbb{B}^N$ is a (K_B, K_I) block-sparse signal over block size d , and $A \in \mathbb{R}^{M \times N}$ is randomly drawn from $\mathcal{A}(R, T, L, d, \alpha)$. Given $y \in \mathbb{R}^M$ in (4) and the block estimate \tilde{x}_ℓ in (11) for block $\ell \in \{1, \dots, L\}$, the probability that $\Gamma(\tilde{x}_\ell)$ in Lemma 2 satisfies for all blocks $\ell \in \{1, \dots, L\}$ is lower bounded by

$$\mathbb{P}(\forall_{\ell \in \{1, \dots, L\}} : \Gamma(\tilde{x}_\ell)) \geq 1 - \delta, \quad (24)$$

if $R = \mathcal{O}(K_B)$ and $T = \mathcal{O}\left(\log\frac{N}{\delta}\right)$, where $0 < \delta < 1$ is an arbitrary target error bound.

1 Proof

Since the block estimate \tilde{x}_ℓ for each block is i.i.d., the probability that $\Gamma(\tilde{x}_\ell)$ in Lemma 2 holds for all $\ell \in \{1, \dots, L\}$ is obtained as

$$\begin{aligned} \mathbb{P}(\forall_{\ell \in \{1, \dots, L\}} : \Gamma(\tilde{x}_\ell)) &= (\mathbb{P}(\Gamma(\tilde{x}_\ell)))^L \\ &\geq \left(1 - \frac{\delta}{L}\right)^L \\ &\geq 1 - L \cdot \frac{\delta}{L} \\ &= 1 - \delta, \end{aligned} \quad (25)$$

where the second inequality follows from Bernoulli's inequality [33] since $\delta/L \ll 1$. \square

Proposition 1 The computational complexity for reliable block support recovery at the BS is of $\mathcal{O}\left(\frac{N}{d} \log N\right)$.

1 Proof

According to (7), for a particular entry x_i of x from block $\ell \in \{1, \dots, L\}$ and a given $t \in \{1, \dots, T\}$, the calculation of its corresponding estimate $\tilde{x}_{t,i}$ only involves a single multiplication. Thereafter, the block-wise estimate \tilde{x}_ℓ in (11) is obtained as the median of $\tilde{x}_{t,i}$ over $\mathcal{O}(Td)$ samples. The computational complexity for finding the median of an unsorted array with N elements is of $\mathcal{O}(N)$ by using the median-of-medians algorithm [35]. Moreover, since the submatrix $A_{t,\ell}$ in (6) has same non-zero entries on the same row, the calculation cost for the block-wise estimate \tilde{x}_ℓ can be further reduced to T times multiplication and $\mathcal{O}(T)$ operations to find the median, resulting in the computational complexity of

⁴ An alternative way of proving this bound is by considering that $\rho \geq \left(1 - \frac{1}{R}\right)^{K_B-1} \geq \theta$ which follows from the derivation in (19). In this case, we obtain again $R \geq \frac{1}{1-\theta^{K_B-1}} \approx \frac{1}{-\frac{1}{K_B-1} \ln \theta} = \frac{K_B-1}{-\ln \theta} = \mathcal{O}(K_B)$, where the approximation follows from the limits of exponential functions since $0 \approx \frac{1}{K_B-1} \ll 1$.

$\mathcal{O}(T)$. To obtain the block-wise estimate \tilde{x}_ℓ for all L blocks and by taking $T = \mathcal{O}(\log N)$ in Lemma 3, the overall computational complexity for reliable block support recovery at the BS is of $\mathcal{O}(TL) = \mathcal{O}(\frac{N}{d} \log N)$. \square

Remark If the cluster size d scales linearly with the increasing network size N , i.e., $d = \mathcal{O}(N)$, the term $\frac{N}{d}$ turns to be an arbitrary constant value. Thus, the algorithm achieves sublinear complexity of $\mathcal{O}(\log N)$ that scales significantly better than conventional approaches.

In short, the above analysis shows that by choosing R and T large enough, i.e., $R = \mathcal{O}(K_B)$ and $T = \mathcal{O}(\log N)$ (for a total of $M = \mathcal{O}(K_B \log N)$ measurements), a reliable block support recovery at the BS can be guaranteed w.h.p. and the computational complexity is of $\mathcal{O}(\frac{N}{d} \log N)$.

4.2 In-block support recovery at devices

Lemma 4 Suppose that $x \in \mathbb{B}^N$ is a (K_B, K_I) block-sparse signal over block size d , and $A \in \mathbb{R}^{M \times N}$ is randomly drawn from $\mathcal{A}(R, T, L, d, \alpha)$. Given $y_D \in \mathbb{R}^M$ in (13) and by applying the algorithm in Sect. 3.3, x_ℓ for block $\ell \in \{1, 2, \dots, L\}$ can be reliably recovered with computational complexity of $\mathcal{O}(dK_I^2)$ if $R = \mathcal{O}(K_B)$ and $T = \mathcal{O}(K_I \log d)$.

1 Proof

As shown in (14), the effective measurements that can be used to perform the in-block support recovery of x_ℓ for a given $\ell \in \{1, \dots, L\}$ comprise only the entries of y_D indexed by $D_\ell \setminus Q_\ell$, where D_ℓ is the index set of rows with non-zero components in $A_{-, \ell}$, and Q_ℓ is the set of collided measurements feedback by the BS. By Lemma 1 and by the fact that the individual estimates are independent, it follows that the overall number of effective measurements T_I for the in-block support recovery of an active cluster $\ell \in \mathcal{S}_B$ can be estimated as

$$T_I = T \cdot \mathbb{P}(\Gamma(\tilde{x}_{t,i})) \geq T \cdot \left(1 - \frac{K_B - 1}{R}\right). \quad (26)$$

To elaborate the in-block support recovery at a device using Algorithm 1, for simplicity we assume that $\tilde{A}_{D,\ell}$ in (14) is real-valued and the system is noise-free. However, the scheme can be easily extended to complex settings and noisy cases as in [36]. Since $\tilde{A}_{D,\ell}$ has independent columns and row-blocks, it follows the column-independent model [37, p. 49] for Subgaussian matrices. Herein $\tilde{A}_{D,\ell}$ can be decomposed as $\tilde{A}_{D,\ell} = \Phi Q$ and we have $y_{D,\ell} = \Phi Q x_\ell = \Phi \tilde{x}_\ell$, where Φ is the column-normalized matrix of $\tilde{A}_{D,\ell}$ and Q is a diagonal matrix with each diagonal entry to be the original norm of the corresponding column. Furthermore, [38] provides a sufficient condition on the measurement matrix for uniform and robust sparse signal recovery, which is the well-known *Restricted isometry property* (RIP). Moreover, according to [37], random matrices with i.i.d. Subgaussian entries and normalized columns have optimal RIP, therefore Φ follows the RIP condition and is admissible for reliable sparse signal recovery. It has been further investigated in

[39] that a K -sparse signal of dimension N can be uniformly reconstructed using OMP with an admissible measurement matrix if its dimension lies in the regime $\mathcal{O}(K \log N)$. While the activation pattern \tilde{x}_ℓ to be reconstructed for cluster ℓ is of dimension d and sparsity level K_I , the support can be recovered w.h.p. if the number of effective measurements T_I satisfies

$$T_I \geq T \cdot \left(1 - \frac{K_B - 1}{R}\right) = \mathcal{O}(K_I \log d). \quad (27)$$

Recall that $1 - \frac{K_B - 1}{R} \geq \theta, \theta \in (\frac{1}{2}, 1)$ should be satisfied for the block support recovery procedure at the BS with $R = \mathcal{O}(K_B)$, then we have $T = \mathcal{O}(\frac{1}{\theta} K_I \log d) = \mathcal{O}(K_I \log d)$ since θ is a given constant.

It is further verified in [40] that the computational complexity is of $\mathcal{O}(NK^2)$ for sparse signal recovery via OMP. Thus, since the signal to be reconstructed for in-block support recovery at a device is of dimension d and sparsity level K_I , $\mathcal{O}(dK_I^2)$ operations are sufficient for decoding with the modified OMP algorithm. \square

4.3 Proof of Theorem 1

For a given realization of the measurement matrix, Lemma 3 guarantees that $R = \mathcal{O}(K_B)$ and $T = \mathcal{O}(\log N)$ are sufficient for reliable block support recovery at the BS, while Lemma 4 shows that $R = \mathcal{O}(K_B)$ and $T = \mathcal{O}(K_I \log d)$ measurements are required for the in-block support recovery at the devices. Taking the maximum value of both cases yields a sufficient condition on the required number of measurements

$$M = \mathcal{O}(\max\{K_B \log N, K_B K_I \log d\}). \quad (28)$$

Furthermore, since the algorithms for block support recovery at the BS requires $\mathcal{O}(\frac{N}{d} \log N)$ operations and the in-block support recovery at the device side requires $\mathcal{O}(dK_I^2)$ operations, the overall computational complexity is of $\mathcal{O}(dK_I^2 + \frac{N}{d} \log N)$ for a successful detection process.

Remark If the signal is treated as a conventional K -sparse vector (where $K = K_B K_I$) as in [11] without exploiting knowledge of the block-sparse structure, a sufficient condition for reliable signal recovery using OMP would be $M = \mathcal{O}(K \log N) = \mathcal{O}(K_B K_I \log N)$ with computational complexity is of $\mathcal{O}(NK^2)$. As M is the length of the unique signature transmitted by an active device, it is also an indication of the signal acquisition time or the communication overhead. Since $d \ll N$ and $K_I \ll K$, we can see that from the scaling point of view, both the communication cost and the computational complexity are significantly reduced by the proposed scheme.

5 Results and discussion

We conduct numerical experiments to verify the performance of the proposed distributed device detection and resource allocation scheme. Mapped into our mathematical model, the target problem is to reconstruct a K -sparse binary vector of length N from M distributed measurements obtained via the measurement matrix $A \in \mathbb{R}^{M \times N}$ which

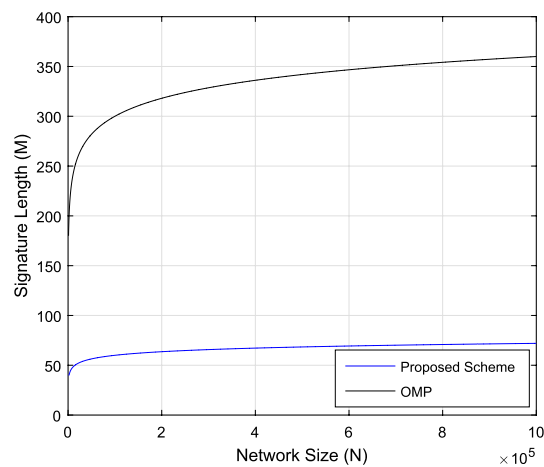
is randomly drawn from $\mathcal{A}(R, T, L, d, \alpha)$ introduced in Sect. 3.1. We compare the performance of the proposed scheme to conventional CS-based approaches, among which we take the classic greedy algorithm OMP with random Gaussian measurements as the baseline. We assume for the baseline that the signal to be reconstructed is treated as a conventional K -sparse vector, and centralized decoding is performed without exploiting knowledge of the block-sparse structure.

In our simulations, we take the number of devices N in the network within the range $[10^3, 10^6]$ and they are partitioned into clusters with equal size $d = 100$. The sparsity level of the signal is set to be $K = 20$, with block sparsity $K_B = 4$ and in-block sparsity $K_I = 5$, respectively. For each plot we average over 1000 pairs of realizations of the measurement matrix and the block-sparse signal.

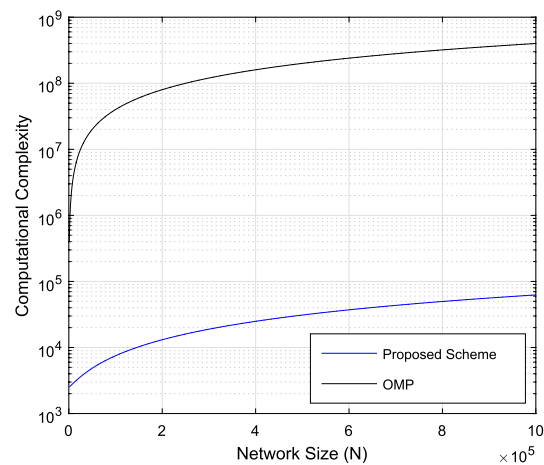
Figure 3a and b give an intuitive comparison between the proposed scheme and standard OMP with Gaussian measurements for reliable signal recovery, in terms of both the length of signatures transmitted by the active devices and the number of operations conducted by the algorithms. It can be seen that the proposed scheme requires a significantly reduced signature length and computational complexity, especially when the number of devices in the network becomes excessively large. As the signature length also implies the signal acquisition time or communication overhead in distributed systems, the proposed scheme leads to a drastically reduced communication cost.

Figure 4a depicts the detection success probability as a function of the signature length for the proposed scheme and the baseline, while taking the network size $N = 10^4$. The performance is evaluated for the noise-free case. We can see that the proposed scheme significantly outperforms standard OMP, where less measurements are required by the proposed scheme to achieve the same detection success probability as the baseline. Figure 4b further extends the evaluation to noisy settings as well as with imperfect channel knowledge. The performance for the noisy case is evaluated by setting the *signal-to-noise ratio* (SNR) to 5 dB in the simulations. In addition, since the channel estimation error can be treated as a component that contributes as an additional source of distortion independent of noise [41], it is modeled as an additive noise term in the measurement matrix with the same variance as the white Gaussian noise. We can see that the proposed scheme achieves significantly higher detection success probability than the baseline under noisy conditions. The performance gain mainly comes from the reliable detection of in-block support cardinality of the active clusters using the sketching algorithm, which sets an appropriate stopping criteria for Algorithm 1 and minimizes the occurrence of false alarms in the detection. Therefore, the proposed scheme shows strong robustness in the presence of noise and imperfect channel knowledge.

We also compare performance of the proposed scheme with two classical random access schemes, namely the LTE RA procedure [5] and the conventional cluster-based approach [42] where a cluster head aggregates messages/requests for the rest of the devices in the cluster and initiates the RA procedure on behalf of the cluster members. We set the number of measurements M to be 839 bits in the simulation—same as the length of Zadoff-Chu sequence [5] used for the LTE RA procedure, thus the three schemes are running with the same signature length. The sparsity level $K = K_B K_I$ is set within the range between 10 and 100.



(a) Signature length for reliable recovery

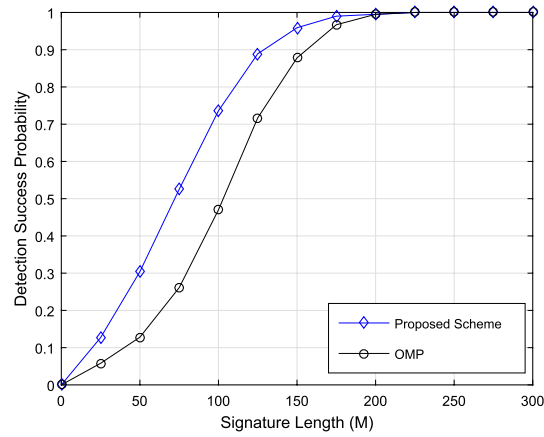


(b) Computational complexity

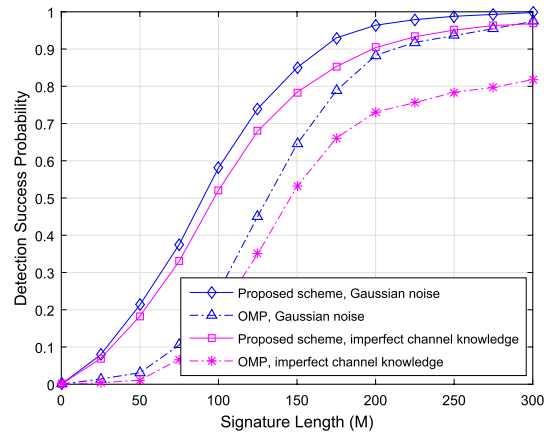
Fig. 3 Performance comparison between the proposed scheme and standard OMP with Gaussian measurements

Figure 5a depicts the detection probability by the three schemes as a function of the number of active devices in the network (i.e., the sparsity level), and Fig. 5b plots the averaged access delay performance of the three schemes. It can be easily observed that the proposed scheme significantly outperforms the LTE RA procedure both in terms of higher detection success probability and reduced access delay, thus achieving much better scalability with the increasing network size and leading to more robustness in the detection process. Moreover, when compared with the cluster-based approach, the proposed scheme also achieves better detection performance if the sparsity level is sufficiently large (≤ 70). Meanwhile, since the proposed distributed scheme is able to avoid the excessive communication and coordination between the devices as well as to the infrastructure as required by the cluster-based approach, the signaling overhead is substantially reduced and thus leading to significantly decreased access latency.

However, there are still some limitations on the proposed approach, especially on the requirement of perfect synchronization during the acquisition phase and priori



(a) Performance in noise-free setting



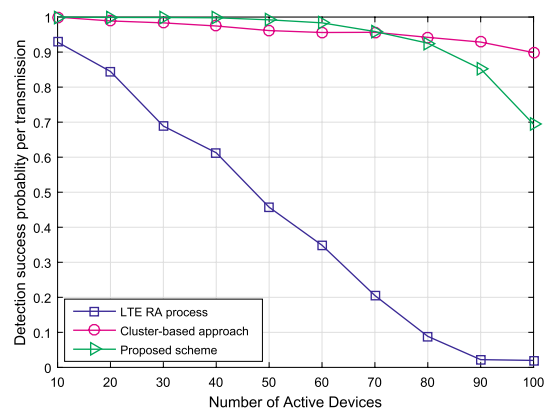
(b) Performance in noisy settings

Fig. 4 Robustness of the proposed scheme compared with standard OMP

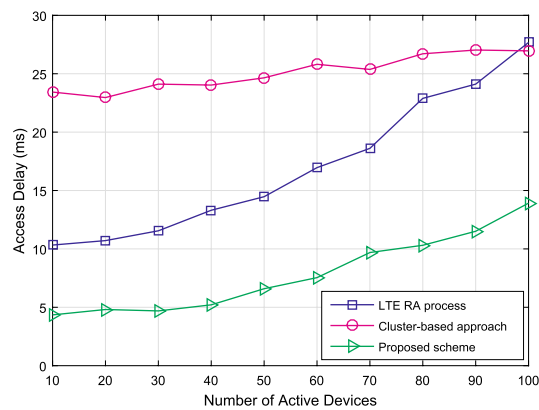
knowledge of the channel state by the M2M devices. Further extensions to relax these limitations will be investigated in future work.

6 Conclusion

This work utilizes the framework of CS for detection of the network activation pattern to facilitate distributed resource allocation in large-scale M2M communication networks. The particular block-sparsity in the activation pattern of the M2M devices is exploited, thus mapping the objective into a support recovery problem for a particular block-sparse signal—with additional in-block structure—in CS based applications. The detection techniques are mainly based on sketching and greedy algorithms, which inherit the virtues of low computational complexity. Furthermore, by applying the distributed ranking-based resource allocation scheme, each active device decides autonomously on which resource to access the channel in a contention-free manner without further coordination, thus excessive control overhead is avoided. It has been verified via theoretical analysis that a (K_B, K_I) block-sparse binary signal $x \in \mathbb{B}^N$ over block size d can be reliably reconstructed using the proposed scheme with $\mathcal{O}(\max\{K_B \log N, K_B K_I \log d\})$



(a) Detection success probability



(b) Averaged access delay

Fig. 5 Performance comparison between the proposed scheme, LTE RA procedure and conventional cluster-based approach [42]

measurements and computational complexity of $\mathcal{O}(dK_l^2 + \frac{N}{d} \log N)$, which achieves a better scaling compared with conventional CS based approaches. Furthermore, the simulation results also reveal the strong robustness of the proposed scheme under noisy conditions and with imperfect channel knowledge.

Abbreviations

M2M	Machine-to-machine
CS	Compressed sensing
IoT	Internet of Things
BS	Base station
OMP	Orthogonal matching pursuit
AMP	Appropriate message passing
RIP	Restricted Isometry property
SNR	Signal-to-noise ratio

Acknowledgements

This work is an extension of [1] presented at the 2016 IEEE International Conferences on Acoustics, Speech and Signal Processing (ICASSP 2016) and is available in part as a preprint [2] on arXiv.

Author contributions

YC contributed to the development of ideas, conducted both theoretical and numerical analysis, and was a major contributor in writing the manuscript. PJ performed the design of the study and provided solid support in the theoretical analysis as well as simulation modeling. CZ participated in the concept design. SS helped draft the manuscript and was responsible for proofreading this work. All authors read and approved the final manuscript.

Funding

This work was partially supported by the DFG in the context of grants JU 2795/2 &3 and STA 864/8-1.

Availability of data and materials

Data sharing is not applicable to this article as no datasets were generated or analyzed during the current study.

Declarations

Competing interests

The authors declare that they have no competing interests.

Received: 24 July 2019 Accepted: 5 July 2022

Published online: 11 September 2022

References

1. Y. Chang, P. Jung, C. Zhou, S. Stanczak, Block compressed sensing based distributed device detection and resource allocation for M2M communications. *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. (2016)
2. Y. Chang, P. Jung, C. Zhou, S. Stanczak, Block compressed sensing based distributed device detection for M2M communications. *arXiv:1609.05080 [cs.IT]*, (2017)
3. D. Boswarthick, O. Elloumi, O. Hersent, *M2M Communications: A System Approach* (John Wiley & Sons Ltd, Hoboken, 2012)
4. H.S. Dhillon, H. Huang, H. Viswanathan, Wide-area wireless communication challenges for the Internet of Things. *IEEE Commun. Mag.* **55**(2), 168–174 (2017)
5. S. Sesia, I. Toufik, M. Baker, *LTE-The UMTS Long Term Evolution, From Theory to Practice*, 2nd edn. (John Wiley & Sons Ltd, Hoboken, 2011)
6. D.A. Schmidt, C. Shi, R.A. Berry, M.L. Honig, W. Utschick, Distributed resource allocation schemes. *IEEE Signal Process. Mag.* **26**, 53–63 (2009)
7. T. Taleb, A. Kunz, Machine type communications in 3GPP networks: potential, challenges and solutions. *IEEE Commun. Mag.* **50**, 178–184 (2012)
8. M. Dohler, D. Boswarthick, J. A. Zarate, Machine-to-machine in smart grids and smart cities. *IEEE Global Telecommunications Conference (GLOBECOM) Workshops*, (2012)
9. E. Candes, J. Romberg, T. Tao, Robust uncertainty principles: exact signal reconstruction from highly incomplete frequency information. *IEEE Trans. Inf. Theory* **52**, 489–509 (2006)
10. E. Candes, M. Wakin, An introduction to compressive sampling. *IEEE Signal Process. Mag.* **25**, 21–30 (2008)
11. D. Donoho, Compressed sensing. *IEEE Trans. Inf. Theory* **52**, 1289–1306 (2006)
12. G. Wunder, H. Boche, T. Strohmer, P. Jung, Sparse signal processing concepts for efficient 5G system design. *IEEE Access* **3**, 195–208 (2015)
13. Y.C. Eldar, P. Kuppinger, H. Bölcskei, Compressed sensing of block-sparse signals: uncertainty relations and efficient recovery. *IEEE Trans. Signal Process.* **58**, 3042–3054 (2010)
14. M. Berlioli, G. Cocco, G. Liva, A. Munari, Modern random access protocols. *Found. Trends Netw.* **10**(4), 317–446 (2016)
15. X. Liu, X. Zhang, Rate and energy efficiency improvements for 5G-based IoT with simultaneous transfer. *IEEE Internet Things J.* **6**(4), 5971–5980 (2019)
16. B. Shim, B. Song, Multiuser detection via compressive sensing. *IEEE Commun. Lett.* **16**, 9725–974 (2012)
17. C. Bockelmann, H.F. Schepker, A. Dekorsy, Compressive sensing based multi-user detection for machine-to-machine communication. *Eur. Trans. Telecommun.* **24**, 1–12 (2013)
18. F. Monsees, M. Woltering, C. Bockelmann, A. Dekorsy, Compressive sensing multi-user detection for multicarrier systems in sporadic machine type communications. *IEEE Vehicular Technology Conference (VTC)*, pp. 1–5, (2015)
19. H. Yin, J. Li, Y. Chai, S.X. Yang, A survey on distributed compressed sensing: theory and applications. *Front. Comput. Sci.* **8**, 893–904 (2014)
20. T. Wimalajeewa, P. K. Varshney, Robust detection of random events with spatially correlated data in wireless sensor networks via distributed compressive sensing. *IEEE International Workshop on Computational Advances in Multi-Sensor Adaptive Processing (CAMSAP)*, (2017)
21. J.A. Tropp, A.C. Gilbert, Signal recovery from random measurements via orthogonal matching pursuit. *IEEE Trans. Inf. Theory* **53**, 4655–4666 (2007)
22. S. Wu, H. Guo, J. Xu, S. Zhu, H. Wang, *In-band full duplex wireless communications and networking for IoT devices: progress, challenges and opportunities* (Elsevier, Amsterdam, 2017). <https://doi.org/10.1016/j.future.2017.10.018>
23. M. Charikar, K. Chen, M. Farach-Colton, Finding frequent items in data streams. in *Proceedings of the 29th International Colloquium on Automata, Languages and Programming*, (2002)
24. J. Haupt, R. Baraniuk, Robust support recovery using sparse compressive sensing matrices. in *Proceedings of the 45th Annual Conference on Information Sciences and Systems*, (2011)
25. M. Goldenbaum, S. Stanczak, Robust analog function computation via wireless multiple-access channels. *IEEE Trans. Commun.* **61**(9), 3863–3877 (2013)
26. Z. Kostić, I. Marić, X. Wang, Fundamentals of dynamic frequency hopping in cellular systems. *IEEE J. Sel. Areas Commun.* **19**(11), 2254–2266 (2001)
27. Z. Zheng, C. Zhu, B. Jiang, W. Zhong, X. Gao, Statistical channel state information acquisition for massive MIMO communications, in *2015 Conference on Wireless Communications and Signal Processing (WCSP)*, (2015)

28. Z. Wang, C. Qian, L. Dai, J. Chen, C. Sun, S. Chen, Location-based channel estimation and pilot assignment for massive MIMO systems. *IEEE International Conference on Communication Workshop (ICCW)*, (2015)
29. F. Kaltenberger, H. Jiang, M. Guillaud, R. Knopp, Relative channel reciprocity calibration in MIMO/TDD systems. 2010 Future Network & Mobile Summit, pp. 1–10, (2010)
30. M. Stege, T. Ruprich, M. Bronzel, G. Fettweis, Channel estimation using long-term spatial channel characteristics, in *International Symposium on Wireless Personal Multimedia Communications (WPMC)*, (2001)
31. L. Liu, W. Yu, Massive connectivity with massive MIMO—part I: device activity detection and channel estimation. *IEEE Trans. Signal Process.* **66**(11), 2933–2946 (2018)
32. O. Knill, *Probability Theory and Stochastic Processes with Applications*, 2nd edn. (Overseas Press, New Delhi, 2009)
33. A.E. Taylor, L'Hospital's rule. *Am. Math. Mon.* **59**, 20–24 (1952)
34. H. Chernoff, A measure of asymptotic efficiency for tests of a hypothesis based on the sum of observations. *Ann. Math. Stat.* **23**, 493–507 (1952)
35. C.E. Leiserson, T.H. Cormen, C. Stein, R. Rivest, *Introduction to Algorithms*, 3rd edn. (MIT Press, UK, 2009)
36. R. Fan, Q. Wan, Y. Liu, H. Chen, X. Zhang, Complex orthogonal matching pursuit and its exact recovery conditions. (2012) [arXiv:1206.2197](https://arxiv.org/abs/1206.2197) [cs.IT]
37. R. Vershynin, Introduction to the non-asymptotic analysis of random matrices. *Compressed Sensing, Theory and Applications*. pp. 210–268, Cambridge University Press, (2012)
38. E. Candes, T. Tao, Decoding by linear programming. *IEEE Trans. Inf. Theory* **51**(12), 4203–4215 (2005)
39. T. Zhang, Sparse recovery with orthogonal matching pursuit under RIP. *IEEE Trans. Inf. Theory* **57**(9), 6215–6221 (2011)
40. J. Wen, Z. Zhou, J. Wang, X. Tang, Q. Mo, A sharp condition for exact support recovery of sparse signals with orthogonal matching pursuit. *2016 IEEE International Symposium on Information Theory*, (2016)
41. J. C. Ikuno, S. Pendl, M. Simko, M. Rupp, Accurate SINR estimation model for system level simulation of LTE networks. *2012 IEEE International Conference on Communications (ICC2012)*, (2012)
42. K. R. Jung, A. Park, S. Lee, Machine-type-communication (MTC) device grouping algorithm for congestion avoidance of MTC oriented LTE network. *Communications in Computer and Information Science*, (2010)