Mobility Support for MIMO-NOMA User Clustering in Next-Generation Wireless Networks

Muhammad Kamran Naeem, *Member, IEEE,* Raouf Abozariba, *Member, IEEE,*
Md Asaduzzaman *Senior Member, IEEE,* and Mohammad Patwary, *Senior Member, IEEE*

Abstract—Non-Orthogonal Multiple Access (NOMA) is a promising technology for future-generation wireless systems, with potential to contribute to the improvement of spectral efficiency. NOMA groups users into clusters, based on channel gain-difference. However, user mobility continuously changes the channel gain, which often requires re-clustering. In this paper, we study a set of re-clustering methods: arbitrary, one-by-one and Kuhn-Munkres assignment algorithm (KMAA), that expedite link re-establishment and keep the clusters interference-free, taking into account the mobility of users. The methods are applied to automatically dissociate identified users within clusters, when the gain-difference is lower than a given threshold, followed by re-association procedure, which integrates users into different clusters, maintaining an appropriate gain-difference. Experimental results show that the KMAA method improves efficiency and capacity through minimizing the number of re-clustering events, improving resource utilization, and lowering signaling overhead. Other sets of results highlight the throughput and outage probability gains of the KMAA method across a wide range of mobility scenarios. We also provide an analysis of the KMAA algorithm when applied to MIMO-NOMA, encompassing link resiliency and maintenance of average gain-difference, among users in clusters.

Index Terms—NOMA, switched beamforming, MIMO-NOMA, NGN, Massive MIMO, 6G.

1 Introduction

Multiple access techniques enable multiple users to share the same bandwidth resource and they are core to cellular communication systems. Previous generations of cellular networks have adopted a range of access methods including Time Division Multiple Access (TDMA) and Orthogonal Frequency Division Multiple Access (OFDMA). These methods are underpinned by the orthogonality in the frequency and time domain in the physical layer, where resource blocks occupied by a user cannot be concurrently shared. Although these techniques facilitated significant advances in terms of data rate and spectrum efficiency in the fourth generation (4G) and 4G Long Term Evolution (LTE) standards of cellular communication, they pose non-trivial challenges to Next Generation Wireless Networks (NGN). The frequency bands currently used in cellular communications are becoming ever more saturated due to high traffic generated by the overwhelming number of smartphones, Internet of Things (IoT) nodes and similar devices [1], [2]. The availability and usage of millimeter-wave frequencies and techniques such as Filter Bank Multi-Carrier Modulation (FBMC) will alleviate the problem, but these are not the only solutions being explored by researchers in industry and academia. On the other hand, Non-Orthogonal Multiple Access (NOMA), a promising technology, aims at improving the spectral efficiency by combining superposition coding at the transmitter with Successive Interference Cancellation (SIC) at the receivers [3], [4], [5]. SIC is a well-known physical layer technique, supports decoding of multiple transmissions at a single receiver. It achieves that by first decoding the strongest signal and treating the rest as interference. It cancels the decoded signal from the ensemble and continues to successively decode the remaining signals (see Fig. 1).

Due to additional system overhead for instantaneous channel feedback requirements and random error propagation, it is not feasible to apply NOMA on all users in a cell as one group. Therefore, user pairing technique has emerged, where NOMA arranges several users into clusters with sufficient gain-difference between channels and assign the same frequency band and time slot [6]. The authors in [7] concluded that the performance gain of Fixed Power NOMA (F-NOMA) over conventional multiple access can only be enlarged by selecting users whose channel conditions are more distinct. It is also mentioned in [8] and [9] that NOMA's performance is determined by how different the users' channel conditions are, allowing users with strong channel to perform SIC.

Wireless channels are characterized by rapid variations of channel quality due to several factors including mobility and location of the communicating devices as well as the movement of objects in space [10], [11]. Even if the initial

- *Muhammad Kamran Naeem is with the School of Engineering, Computing and Mathematical Sciences, University of Wolverhampton, Wolverhampton WV1 1LY, UK.*
  E-mail: kamran.naeem@ieee.org
- *Raouf Abozariba is with the School of Computing and Digital Technologies, Birmingham City University, Birmingham B4 7XG, UK.*
  E-mail: r.abozariba@ieee.org
- *Md Asaduzzaman is with the School of Creative Arts and Engineering, Staffordshire University, Stoke-on-Trent ST4 2DE, UK.*
  E-mail: asad@ieee.org
- *Mohammad Patwary is with the School of Engineering, Computing and Mathematical Sciences, University of Wolverhampton, Wolverhampton WV1 1LY, UK.*
  E-mail: m.n.patwary@ieee.org

Manuscript received Month Day, Year; revised Month Day, Year.

NOMA user pairing/clustering is performed optimally, the inherent high mobility characteristics of users in cellular networks can rapidly influence the channel gain-difference, raising the possibility of users being paired incorrectly and the sum throughput may drop significantly. In this paper, we perform continuous per-cluster analysis to determine the validity and effectiveness of each cluster by monitoring the instantaneous channel gain-difference between users. For cases where a cluster is found to be ineffective, we developed a dissociation mechanism, where one or more users are removed from the cluster to ensure SIC is successfully achieved for the remaining users in the cluster.

Removing users from a cluster means they will be unable to use the bandwidth configurations associated with the original cluster. To continue the operation of NOMA, the dissociated users should swiftly merge with a new suitable host cluster and obtain new network configurations. However, in congested urban networks, the number of clusters could be large and there could be numerous choices for each user, but only a few are optimal. In this paper we compare three different clustering methods and analyze their effectiveness from multiple new parameters.

In multiple-input multiple-output (MIMO) NOMA where users in a cluster also share the same beam, the issue of maintaining the gain difference under mobility is even more complicated compared to a 360^o omnidirectional mode [12]. While the narrowest beamwidth sectors offer maximum directivity gain [13], in the presence of mobility, it is difficult to maintain users under acceptable gain difference. To solve this problem, we deploy our highest performance clustering method to maintain cluster order.

Our contributions are summarized as follows:

- A mobility management scheme for NOMA, which explicitly guarantees the gain-difference in all clusters at a desired level, achieved through dissociation and re-association mechanisms. In particular, our solution takes into account high user mobility in dense urban environments.
- Three different re-association methods: arbitrary, one-by-one and Kuhn-Munkres assignment algorithm (KMAA), performed in real-time to address the shortcomings of power allocation techniques. The KMAA algorithm is powered by the dynamic Hungarian assignment algorithm, which is proposed to improve the re-association accuracy while maintaining the run-time efficiency.
- Detailed theoretical analysis and simulation results examine the performance of our proposed methods and provide extensive comparisons, highlighting the advantageous of the KMAA algorithm, minimizing outages and maximizing throughput in a robust manner.
- Analysis of link resiliency and average gain difference for MIMO-NOMA beamforming with varied beam-widths, under high mobility. Beam and cluster re-selections are achieved by deploying the KMAA algorithm.

The proposed solutions are useful when the physical layer properties are rapidly changing, requiring frequent instantaneous re-clustering processes. This paper is a continuation of our previous work in [14] and here, an attempt is made to derive the outage probability, taking into account the proposed dissociation and re-association mechanisms. We also added clustering comparison using Jaccard Coefficient, a method which was not used previously in the this context. Furthermore, we derived more results on the impact of mobility intensity on the proposed dissociation and reassociation and thoroughly investigated the KMAA algorithm with beam-switching under MIMO-NOMA set up.

The rest of the paper is organized as follows. Section 2 contains the relevant background information. Section 3 presents the system model. In Section 4 the problem is described mathematically followed by the resource allocation procedures in Section 5. We extend our system model from SISO to MIMO-NOMA and we discuss the associated challenges in Section 6. The simulation results are presented in Section 7 and we conclude with a few remarks in Section 8. Table 1 is a summary of key symbols used throughout this paper.

2 Related work

In this section we briefly review a number of studies which address clustering problems in NOMA. In [15], the gain difference is artificially created using precoding and detection strategies. Particularly, the precoding matrix is designed to degrade user's (nearer to the base station) effective channel gains while improving the signal strength for the user farther from the base station. Complex power allocation strategies such as cognitive radio power allocation was also proposed in [7]. However, in practice, power allocation techniques fail to create a gain-difference when users of one NOMA cluster mutually move closer to the edge of the cell, where the signal severely diminishes, and the residual noise accumulates. In this case, power domain multiplexing is ineffective in enabling SIC, since lowering the transmit power for users at the edge will result in the received power being near the noise floor, making signal decoding impossible. This will not only result in low cluster throughput but also unnecessary waste of wireless resources, including power and spectrum. In addition, the majority of power-based user pairing algorithms in the literature are designed based on exhaustive search methods, which are computationally complex [16], hence limiting the number of users per cluster. Other solutions used learning approaches to facilitate the association between users and BS in an iterative and distributed manner [17]. In contrast, our proposed system is centralized in nature following the cellular communication design principles.

The authors in [18] and [19] proposed maximum weighted bi-partite matching and sequential maximum weighted multi-partite matching algorithms to perform clustering, respectively. However, these studies use drop-based scenarios with limited mobility consideration and do not address the re-clustering problems caused by user movement or channel variation over time. [20] developed a dynamic K-means online based machine learning algorithm to account for continuous arrival of users. This study, however, failed to take account of the intra-cell mobility and its impact on clusters' validity. In large-scale NOMA systems, it is hard to find an optimal solution to the user

clustering problem. Moreover, to the best of our knowledge, the published works have not thoroughly considered the impact of mobility in respect to the validity of NOMA clusters, in dense mobile environment. Therefore, it remains unclear whether the available solutions are effective for highly mobile scenarios.

In addition, the effectiveness of these clustering techniques is not analyzed under MIMO beamforming. Notable studies that combined NOMA with beamforming are [12] and [21] where heuristic clustering methods are developed. Greedy clustering algorithm were proposed in [22] to combat inter-cluster and intra-cluster interference. An enhanced K-means scheme is proposed in [23] to reduce the interference inter-beams. However, the study addressed the maximization of energy efficiency problem in THz band. While these clustering methods provide high throughput, they do not come with mobility support, nor discuss the effect of different antenna beamwidths. In this paper, we integrate our most advanced solution with beam switching in large-scale MIMO-MOMA system to study the link resiliency and the impact on the average channel gain-difference, under various beamwidths.

Furthermore, the majority of previous work on NOMA focused on the sum rate and outage probabilities optimization by using different power allocation strategies. In this paper, we argue that mobility should be taken into account when analyzing solutions targeted for NOMA technology.

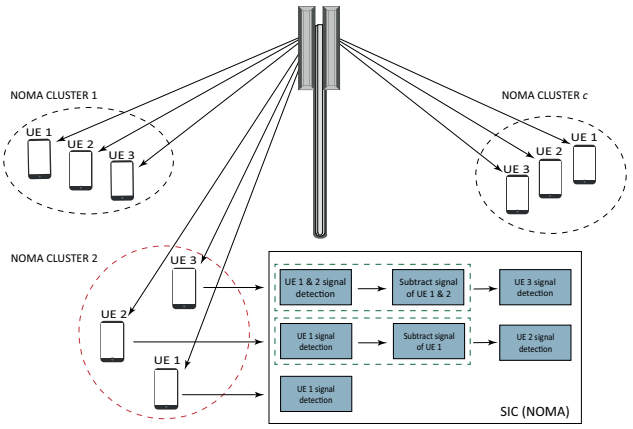

Fig. 1. A downlink NOMA system with multiple clusters

3 System Model

Consider a Base Station (BS) supporting $|\mathcal{V}|$ User Equipment (UEs), randomly placed across the cell. The UEs are considered to be mobile or stationary and their locations may change with respect to the BS over time. The magnitude and angle of mobility is determined by Random Waypoint mobility model, widely used in the simulation studies of cellular networks [24]. We assume that there is a set of users who are mobile, and they move between different cells, capturing the real-world dynamics in cellular networks.

In this model, we assume that the BS has access to the complete set of data on UEs, including two dimensional positions and the channel coefficients of each UE. Consider $\mathcal{V}'$ is a set of all users under coverage and mathematically this can be defined by $\mathcal{V}' = \{v_1, v_2, \ldots, v_i, \ldots, v_{|\mathcal{V}'|}\}$. To group UEs into clusters, we use the distance-based criteria (Euclidean distance), for which a distance measure is specified between each pair of UEs, in respect to the BS. The BS forms the clusters $\mathbf{c}_i$ based on criteria relevant to the UEs distances $\mathcal{D} = \{d_1, d_2, \ldots, d_i, \ldots, d_{|\mathcal{V}'|}\}$ from each other.

TABLE 1
Key symbols and definitions

Symbols & Definitions	
$\mathcal{V}$	set of users in the network
$\mathcal{V}'$	set of users under coverage
k_{max}	maximum number of users in a cluster
$\mathcal{C}$	set of clusters in the network
$\mathcal{C}^*$	set of clusters updated after mobility
$\mathcal{R}$	available resource blocks
$\mathcal{M}$	set of dissociated users
$\mathcal{N}$	set of under-utilized clusters
h_i	channel gain of ith user
B	size of the resource block
P_i	transmit power for ith user
R	total capacity of the cell

In the initial clustering, UEs are first arranged in ascending order based on $\mathcal{D}$ (described in more details in Section 5). The set of available resources at each BS is $\mathcal{R}$, with $|\mathcal{R}| = N_{RB}$, where each resource block (RB) represents the minimum spatio-temporal scheduling unit. The number of RBs available bounds the number of clusters in the network. Let l be the number of UEs in each group, where

$$l = \left\lceil |\mathcal{V}'|/k_{max} \right\rceil, \tag{1}$$

and k_{max} is the maximum number of users allowed in each cluster. The groups of UEs can be arranged in $\mathbf{C}$, which is defined as

$$\mathbf{C} = \begin{bmatrix} \mathbf{c}_1 & \mathbf{c}_2 & \ldots & \mathbf{c}_k \\ v_1 & v_2 & \ldots & v_l \\ v_{l+1} & v_{l+2} & \ldots & v_{2*l} \\ \vdots & \vdots & \vdots & \vdots \\ v_{l(k-1)+1} & v_{l(k-1)+2} & \ldots & v_{k*l} \end{bmatrix}, \tag{2}$$

where each column of $\mathbf{C}$ represents the UEs in a cluster. Consider $\mathcal{C}$ is the initial set of clusters of size $(1 \leq \mid \mathbf{c}_i \mid \leq k_{\max})$ in the network which is defined as

$$\mathcal{C} = \{\mathbf{c}_1, \mathbf{c}_2, \ldots, \mathbf{c}_i, \ldots, \mathbf{c}_k\}, \tag{3}$$

where k represents the total number of clusters. Moreover, the sum of all users within the BS is denoted as

$$\mathbf{c}_1 \bigcup \mathbf{c}_2 \bigcup \ldots \bigcup \mathbf{c}_k \equiv \{1, 2, \ldots, V'\}$$

where

$$\mathbf{c}_p \bigcap \mathbf{c}_q \equiv \emptyset, \quad p \neq q.$$

The size of the clusters $k_{\max}$ are defined by the range of the BS coverage and throughput requirements. The number of available resource blocks $|\mathcal{R}|$ is also taken into consideration when deciding the cluster size. For example the following relation can be used

$$k_{max} = \frac{|V'|}{|\mathcal{R}|}. \tag{4}$$

Note that in this paper, we assumed that the base station owns a sufficient amount of bandwidth resources, as our focus is more on the clustering, and not so much on resource blocks management. Our ultimate goal is to minimize the number of clusters, which will have an indirect impact on the required number of resource blocks. However, the treatment of resource blocks would not add any complications to the system. Resource block management is performed through Radio Resource Control, which is an independent entity within the cellular networking stack. Upon detection of low channel gain-difference, the UE, which is likely to cause the highest disruption to the overall cluster, is dissociated from their clusters. The process of UE dissociation and re-association is presented in the following section.

4 Problem Setup

NOMA arranges the users into clusters, and to keep the clusters with high gain-difference and to enable users perform SIC successfully, our method discards UEs which reduces channel gain difference in the cluster below a specified threshold. This process is defined as dissociation. The method also provides the identity of vacant positions, which allow new and dissociated UEs to join new clusters. This joining process is defined as re-association. The dissociation and re-association processes are simplified in Fig. 2.

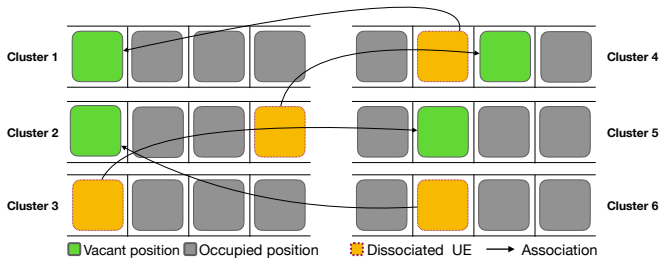

Fig. 2. An example of dissociation and re-association in shared clusters

4.1 Dissociation

In real world environments, wireless transmission suffers from various channel impairments, including path-loss, shadowing, and fading. The channel of a wireless signal traversing a multiple N path is usually represented as linear combinations of complex exponential, given by [25]

$$h = a \exp\left(-j \cdot 2\pi \frac{d}{\lambda} + j\phi\right), \tag{5}$$

where λ is the wavelength, a is the path attenuation, d is the distance the path traverses, and ϕ is a frequency-independent phase that captures whether the path is direct or reflected. As the signal travels through N paths, the channel at the receiver can be expressed as:

$$h = \sum_{n=1}^{N} a_n \exp\left(-j \cdot 2\pi \frac{d_n}{\lambda} + j\phi_n\right). \tag{6}$$

The N paths represent multi-path fading, which are uniformly distributed and mutually independent random variables. The distribution ensures all frequency components are affected almost equally. Multi-path fading is a type of small-scale fading that is dominant in mobility [26].

Now, consider any two channel gain measurements h_x^c, h_y^c, for the channel between the BS and any two UEs x and y in the cluster $\mathbf{c} \in \mathcal{C}$, respectively. The UEs in cluster $\mathbf{c}$ share the same channel on NOMA basis. As mentioned in Section 1, the performance gain of NOMA increases in channel gain, i.e., when difference in path-loss between any set of UEs in one cluster is large. We define the dissociation procedure employed by the BS as

$$\mathcal{D}(h_x^c, h_y^c) = \begin{cases} 1, & \text{if } h_x^c \text{ and } h_y^c \text{ satisfy a certain condition,} \\ 0, & \text{otherwise.} \end{cases}$$

Decisions to disassociate UEs from clusters is based on binary decision rule $\mathcal{D}(h_x^c, h_y^c)$, which is founded on a chosen distance between h_x^c and h_y^c and determined by the following model

$$\mathcal{D}\left(h_x^c, h_y^c\right) = \mathbf{1}\left\{\|h_x^c - h_y^c\| < \lambda_0^c\right\}, \tag{7}$$

for each $x, y \in \mathbf{c}$ and $x \neq y$, where $\mathbf{1}$ denotes the indicator function that takes the value 1 if its argument is true and 0 otherwise. To maintain the cluster optimality in the system, we choose one of the UEs in the cluster to be dissociated using the rule

$$\begin{aligned} (\mathcal{P}_1) \quad \max \quad & \sum_{\mathbf{c}_1, \mathbf{c}_2, \ldots, \mathbf{c}_k} \sum_{x_i, y_i \in \mathbf{c}_i} D(h_{x_i}^c, h_{y_i}^c)\, Z_{x_i, y_i} \\ \text{s.t.} \quad & D(h_{x_i}^c, h_{y_i}^c) \leq \lambda_0^c, \ \forall\, c \in \mathcal{C}, \\ & Z_{x_i, y_i} \in \{0, 1\}, \end{aligned} \tag{8}$$

where

$$D(h_{x_i}^c, h_{y_i}^c) = \|h_x^c - h_y^c\|, \tag{9}$$

and λ_0^c is the chosen threshold such that $0 < \lambda_0^c < \lambda'$, where $\{\lambda_0^c\}_{c \in \mathcal{C}}$ are the thresholds relative to the cluster size and λ' relative to the cell size, describing the desired degree of channel gain-difference where large λ_0^c gives less channel gain-difference. The optimal threshold can be based on several factors including application requirements, channel bandwidth availability, slow/fast fading, number of users in a cluster, mobility intensity and noise in the channel. One can find the optimal threshold value either by solving a multi-objective optimization problem or by conducting empirical analysis. This paper does not attempt to solve this general problem.

It is important to note that the BS needs to observe the values of h_x^c and h_y^c to make an informed decision.

- If $\mathcal{D}\left(h_x^c, h_y^c\right) = 1$, this implies that observations of channel gain difference from users x and y, in the same cluster are too close, therefore one of the UEs (x or y) is dissociated.
- If $\mathcal{D}\left(h_x^c, h_y^c\right) = 0$, then we do not consider the difference in gain to be close enough to impact the SIC process.

Dissociating UEs can simply be achieved by issuing a disassociate request to the UE.

4.2 Re-association

Dissociated UEs can switch to operate under OMA scheme, requiring additional resources. This is attainable if there are additional channels reserved for such circumstances and may provide low latency. However, this is undesirable practice, when channels are a scarce commodity, and we only consider this option when a suitable cluster cannot be found. The proposed dissociation/re-association method balances achieving high throughput with fairness as it ensures all users remain linked to the best possible cluster at any given time. To reduce computational overhead, instead of computing throughput for every possible combination of clusters, we just perform dissociation and re-association for a subset of users. Due to the dynamic nature of cellular networks, dissociated UEs may join one of the clusters with empty positions which are given by

$$\sum_{i=1}^{|\mathcal{C}|}\sum_{j=1}^{|\mathbf{c}|}\mathcal{D}_{ij}\left(h_x^c, h_y^c\right) = m, \tag{10}$$

where $\mathcal{D}_{ij}(\cdot)$ is the decision value $\{0,1\}$ for jth user in the ith cluster, m represents the total number of vacant positions in the network. In NOMA, it is desired to maximize the distance between UEs in a cluster, for example, in a cluster of two UEs the optimal clustering is in

$$\begin{aligned} &\underset{\substack{x_i,x_j\in\mathbf{c}\\ i\neq j}}{\arg\max} \quad \mathcal{D}(h_{x_i}^c, h_{x_j}^c) \\ &\text{subject to} \quad \|h_{x_i} - h_{x_j}\| > \lambda_1^c \;\; \forall\, i,j \in \mathbf{c}. \end{aligned} \tag{11}$$

Now we shall discuss the criteria of joining users into a cluster. Let $\mathbf{X} = \{x_{i1}, x_{i2}, \ldots, x_{in_i}\}$ be a set of n_i UEs dissociated from the ith cluster $\mathbf{c}_i$, $\mathbf{c}_i \in \mathcal{C} = \{\mathbf{c}_1, \mathbf{c}_2, \ldots, \mathbf{c}_k\}$. The set of remaining UEs in $\mathcal{C}$ forms a new cluster $\mathcal{C}^*$ with $n = \sum_{i=1}^{k} n_i$ vacant positions such that $\mathcal{C}^* = \{\mathcal{C} \setminus \mathbf{X}\}$. Let Z_{x_i,x_j} be a binary indicator variable, which takes value 1 if the ith element of jth cluster $Z_{x_i,x_j} \in \mathbf{c}_i$ is assigned to the new cluster $\mathbf{c}_i^*$ i.e., $Z_{x_i,x_j} \in \mathcal{C}$ is assigned to $\mathcal{C}^*$. Therefore, the new set of cluster can be represented by $\mathcal{C}^* = \{\mathbf{c}_1^*, \mathbf{c}_2^*, \ldots, \mathbf{c}_k^*\}$. The optimization problem of UE association with the appropriate constraints can now be written as follows

$$\begin{aligned} (\mathcal{P}_2) \quad \max \quad & \sum_{\mathbf{c}_1,\mathbf{c}_2,\ldots,\mathbf{c}_k^*}\;\sum_{x_i,x_j\in\mathbf{c}_i,\mathbf{c}_i^*} D(h_{x_i}^c, h_{x_j}^c)\, Z_{x_i,x_j} \\ \text{s.t.} \quad & D(h_{x_i}^c, h_{x_j}^c) > \lambda_1^c \;\; \forall\, c \in \mathcal{C}, \\ & \mathbf{c}_i^* \bigcap \mathbf{c}_j^* = \phi \;\; \forall\, i,j \in \mathcal{C}^*, \\ & Z_{x_i,x_j} \in \{0,1\}, \end{aligned} \tag{12}$$

where

$$D(h_{x_i}^c, h_{y_i}^c) = \|h_x^c - h_y^c\|, \tag{13}$$

h_{x_i} is the channel gain between the BS and the dissociated user x_i and h_{y_j} is the channel gain between the BS and an element of the $\mathbf{c}^*$. The re-association problem $(\mathcal{P}_2)$ is a binary optimization problem and can be viewed as a clustering technique. The optimization problem can be solved using three different approaches: (i) associating users arbitrarily, which does not guarantee the optimality (ii) associating users one-by-one, which would give near optimal solution but does not guarantee the global optimal and (iii) associating users simultaneously, which guarantees the global optimal solution.

The arbitrary solution solves the problem through assigning each dissociated user to a suitable cluster under a certain gain-difference threshold. The one-by-one solution assigns dissociated users to the best cluster available from among all possible cases, to reduce the interference, under a certain gain-difference threshold. In the KMAA solution we deploy the Hungarian algorithm [27], [28] (also known as the Kuhn-Munkres algorithm) maximizing $D(h_x^c, h_y^c)$. These solutions are described next.

5 NOMA RESOURCE ALLOCATION THROUGH USER DISSOCIATION/RE-ASSOCIATION PROCEDURE

We present here the solution for the re-clustering problem ($\mathcal{P}_1$ & $\mathcal{P}_2$), where Algorithm 1, phase 1 is an initial clustering, based on Euclidean distance, that is being called at the start of running phase 2.

Overview of Algorithm 1 — Phase 1: The algorithm computes an approximate solution based on distance from the base station in order to initialize the maximization of the objective function of $\mathcal{P}_2$. Phase 1 performs the initial clustering and is composed of the following steps: **Step 1:** Construct the set $\mathcal{D}$, which consists of the distances between each UE and the BS and sort all the UEs, based on their distance from BS (Line 4–7). **Step 2:** Construct k groups containing UE from $\mathcal{V}'$, where each group consists of l number of UEs. We assign the first l UEs to form group 1, then the next set of l UEs to form group 2 and so on. The first (second) user in each group will form the first (second) cluster and so on (Line 8–13).

NOMA will benefit from the location and positioning enhancements which are new features for wireless standards [29]. Future studies will also define more accurate sets of positioning techniques for both indoor and outdoor environments. The form of clustering in Algorithm 1 (Phase 1) is the de facto method in NOMA allocation and used by several researchers [30]. However, this user clustering technique would inevitably be invalid in networks where users are mobile. Our next algorithms deal with this problem. The detailed pseudo-code is presented in Algorithm 1.

Phase 2: In phase 2 of Algorithm 1, we applied a backtracking procedure to continuously monitor and update the clusters. The pseudo-code consists of the following steps: **Step 1:** Monitor the gain-difference between each pair of UEs within a cluster in real-time (Line 15–20). **Step 2:** If the gain-difference between any two UEs within a cluster falls below λ_1 but remain greater than λ_0, then the UEs in that cluster are prompted to increase the frequency of obtaining and reporting channel gain measurements. This minimizes channel feedback and communications overhead, when user mobility in the network is low (Line 21–23). **Step 3:** If the gain-difference between UEs within a cluster is less than the given threshold λ_0, the UE which is introducing interference to other UEs within the cluster is dissociated (Line 24–30). **Step 4:** Assign the dissociated users to set $\mathcal{M}$ and save the cluster indexes from which the users are dissociated in $\mathcal{N}$.

Algorithm 1: Initial clustering and dissociation procedure

1 **Phase 1: Choosing initial clusters**
2 **Input:** Number of UEs $|\mathcal{V}'|$, cluster size $k_{\max}$ ▷ e.g., $(1 \leq |K| \leq k_{\max})$, λ_0 and λ_1 gain-difference threshold values, $\mathcal{N}$ is a set of available positions and $\mathcal{M}$ is a set of dissociated and new UEs
3 **Output:** $\mathcal{C}$;
4 **for** $i \in \{1, 2, \ldots, |\mathcal{V}'|\}$ **do**
5 calculate the distance between UE v_i and BS and save it in set $\mathcal{D}$
6 sort $\{\mathcal{D}\}$
7 sort $\{\mathcal{V}'\}$ according to $\{\mathcal{D}\}$
8 **for** $j \in \{1, 2, \ldots, k_{\max}\}$ **do**
9 $l = j \times k$
10 **if** $l \leq |\mathcal{V}'|$ **then**
11 $\mathbf{c}_j = \{(k \times (j-1)) + 1, (k \times (j-1)) + 2, \ldots, (k \times (j-1)) + k_{\max}\}$ assigning UEs to the clusters $\{1, 2, \ldots, k\}$.
12 **else**
13 $\mathbf{c}_j = \{(k \times (j-1)) + 1, (k \times (j-1)) + 2, \ldots, |\mathcal{V}'|\}$ assigning UEs to the clusters $\{1, 2, \ldots, |\mathcal{V}'| - (k \times (j-1))\}$.
14 **Phase 2: Managing clusters**
15 **for** $k = 1 \rightarrow |\mathcal{C}|$ **do**
16 $\mathbf{s} \leftarrow \mathbf{s}_{emp}$ ▷ an empty array.
17 $\mathbf{W} \leftarrow ((1 : (|\mathbf{c}_k| - 1)) \times (1 : |\mathbf{c}_k|))$ ▷ is a matrix of zeros, where each υ represents the status of UE interference.
18 **for** $i = 1 \rightarrow (|\mathbf{c}_k| - 1)$ **do**
19 **for** $j = i + 1 \rightarrow |\mathbf{c}_k|$ **do**
20 $d_{ij} = |(h_{x_i} - h_{y_j})|$ ▷ d_{ij} is a function to find the gain-difference between all the UEs in a cluster.
21 **if** $\lambda_0 \leq d_{ij} \leq \lambda_1$ *(where* $\lambda_0 < \lambda_1$*)* **then**
22 $\mathbf{c}_k \leftarrow \mathbf{c}_k$
23 increase the rate of channel feedback measurements.
24 **else if** $d_{ij} \leq \lambda_0$ **then**
25 $\mathbf{s} \leftarrow [\mathbf{s}, \upsilon_{(i,j)}]$ ▷ save index of interfering UEs in $\mathbf{s}$.
26 $\mathbf{W}(i, j) \leftarrow 1$ ▷ interference status of UEs becomes 1 when their gain-difference is less than λ_0.
27 **for** $l = 1 \rightarrow (1 : |\mathbf{c}_k|)$ **do**
28 $\psi \leftarrow \mathbf{W}(:,\ l)$ ▷ identify the UEs which are causing interference by analyzing each column of $\mathbf{W}$.
29 $\mathbf{c}_k \leftarrow (\mathbf{c}_k \setminus \psi)$ ▷ disassociate UE with desired criteria
30 check $\mathbf{s}$ if all the interfering UEs are dissociated from k^{th} cluster otherwise go to 27.
31 $\mathcal{M} \leftarrow \psi$ ▷ set of all the dissociated UEs
32 $\mathcal{N} \leftarrow$ save cluster number k
33 **if** $|\mathbf{c}_k| < k_{max}$ **then**
34 $\mathcal{N} \leftarrow$ save cluster number k

Find all the clusters which are under-utilized and save their indexes in $\mathcal{N}$ (Line 31–34).

The dissociated UEs from the set $\mathcal{M}$ are then assigned to clusters in set $\mathcal{N}$, using arbitrary, one-by-one or the KMAA algorithms, following the corresponding techniques therein. The dissociated users are assigned to new suitable positions in other clusters, within multiple position candidates. It is reasonable for the UEs to be placed to the position where the gain-difference between the users in the new cluster is higher than λ_1. The logic behind this algorithm is to keep track of the gain-difference changes between users which occur as a consequence of mobility and to maintain the clusters in working order, minimizing outages. The detailed pseudo-code is presented in phase 2 of Algorithm 1.

Algorithm 2: Re-association algorithm using arbitrary mechanism

1 **Input:** Let λ_0 gain-difference threshold value, $\mathcal{N}$ is a set of available positions and $\mathcal{M}$ is a set of dissociated and new UEs.
2 $\mathcal{M}^* \leftarrow \emptyset$
3 **for** $i = 1, 2, \ldots, |\mathcal{M}|$ **do**
4 $\rho \leftarrow 0$
5 **for** $j = 1, 2, \ldots, |\mathcal{N}|,\quad j \neq i$ **do**
6 ▷ So we will go through all the clusters except $\mathbf{c}_i$ from which ith user is dissociated
7 $\psi \leftarrow (1 \times |K_j|)$ is a vector of zeros
8 **for** $l = 1, 2, \ldots, |K_j|$, **do**
9 $d \leftarrow |(h_{x_i} - h_{x_l})|$
10 **if** $d \geq \lambda_0$ **then**
11 $\psi_l = 1$
12 **if** $\sum \psi = |K_j|$ **then**
13 $\mathbf{c}_j \leftarrow \mathbf{c}_j \bigcup m_i$
14 **if** $|\mathbf{c}_j| = k_{max}$ **then**
15 $\mathcal{N} \leftarrow \mathcal{N} \setminus \mathbf{c}_j$
16 $\rho \leftarrow 1$
17 **continue**
18 **if** $\rho = 0$ **then**
19 $\mathcal{M}^* \leftarrow \mathcal{M}^* \bigcup m_i$
20 Add new clusters for UEs in $\mathcal{M}^*$ using Algorithm 1 from Line (4–13)

Overview of Algorithms 2, 3 and 4 — Following the scanning procedure to identify vacant positions, Algorithm 2 performs assigning UE to clusters randomly (Line 5-13) and update all the clusters accordingly. Algorithm 3 performs a search to find the maximum gain difference between users in clusters containing valid positions (Line 5-16) and update all the clusters accordingly.

To optimize the assignment problem and to maintain the global maximum-gain difference, Algorithm 4 performs the best possible solution, which is achieved using the Hungarian algorithm in two steps: (1) Line 4-17, we construct matrix, $\mathbf{G}$, which contains channel gain-difference values between the dissociated UE in $\mathcal{M}$ and UE in $\mathcal{N}$. (2) The assignment of UE to clusters is determined by Line 18-31 and saved in $\mathbf{G}^*$, followed by updating the sets $\mathcal{M}$ and $\mathcal{N}$ (Line 32-37).

Algorithm 3: Re-association algorithm using one-by-one mechanism

Input: Let λ_0 gain-difference threshold value, $\mathcal{N}$ is a set of available positions and $\mathcal{M}$ is a set of dissociated and new UEs.
$\mathcal{M}^* \leftarrow \emptyset$
for $i = 1, 2, \ldots, |\mathcal{M}|$ **do**
  $\rho \leftarrow 0$, $\mathbf{d}' \leftarrow (|\mathcal{N}| \times 1)$ is an array of zeros,
  for $j = 1, 2, \ldots, |\mathcal{N}|, \quad j \neq i$ **do**
    ▷ So we will go through all the clusters except $\mathbf{c}_i$ from which ith user is dissociated
    $\psi \leftarrow (1 \times |K_j|)$ is a vector of zeros
    $\mathbf{d} \leftarrow (1 \times |K_j|)$ is a vector of zeros
    for $l = 1, 2, \ldots, |K_j|$, **do**
      $d \leftarrow |(h_{x_i} - h_{x_l})|$
      if $d \geq \lambda_0$ **then**
        $\psi_l \leftarrow 1$
        $\mathbf{d}_l \leftarrow d$
    if $\sum \psi = |K_j|$ **then**
      $\mathbf{d}'_j \leftarrow \sum \mathbf{d}_l$
  $\mu \leftarrow \max(\mathbf{d}')$ and save its position to v
  re-associate the UE m_i with cluster $\mathbf{c}_v$
  $\mathbf{c}_v \leftarrow \mathbf{c}_v \bigcup m_i$
  $\rho \leftarrow 1$
  if $|\mathbf{c}_v| = k_{max}$ **then**
    $\mathcal{N} \leftarrow \mathcal{N} \setminus \mathbf{c}_v$
  if $\rho = 0$ **then**
    $\mathcal{M}^* \leftarrow \mathcal{M}^* \bigcup m_i$
Add new clusters for UEs in $\mathcal{M}^*$ using Algorithm 1 from Line (4–13)

Algorithm 4: Re-association algorithm using simultaneous mechanism

Input: Let λ_0 gain-difference threshold value, $\mathcal{N}$ is a set of available positions and $\mathcal{M}$ is a set of dissociated and new UEs.
$\mathcal{M}^* \leftarrow \emptyset$
$\mathbf{G} \leftarrow (|\mathcal{M}| \times |\mathcal{N}|)$ is an array of zeros,
for $i = 1, 2, \ldots, |\mathcal{M}|$ **do**
  $\rho \leftarrow 0$, $\mathbf{d}' \leftarrow (|\mathcal{N}| \times 1)$ is an array of zeros,
  for $j = 1, 2, \ldots, |\mathcal{N}|, \quad j \neq i$ **do**
    ▷ So we will go through all the clusters except $\mathbf{c}_i$ from which ith user is dissociated
    $\psi \leftarrow (1 \times |K_j|)$ is a vector of zeros
    $\mathbf{d} \leftarrow (1 \times |K_j|)$ is a vector of zeros
    for $l = 1, 2, \ldots, |K_j|$ **do**
      $d \leftarrow |(h_{x_i} - h_{x_l})|$
      if $d \geq \lambda_0$ **then**
        $\psi_l \leftarrow 1$
        $\mathbf{d}_l \leftarrow d$
    if $\sum \psi = |K_j|$ **then**
      $\mathbf{g}_j \leftarrow \sum \mathbf{d}_l$
  $\mathbf{G} \leftarrow \mathbf{g}$
$n \leftarrow$ number of rows in $\mathbf{G}$
rowMin$[i] \leftarrow i$
colMin$[j] \leftarrow j$
assignMat $\leftarrow \mathbf{G}^*$
while $Num < n$ **do**
  for $i = 1, 2, \ldots, n$ **do**
    for $j = 1, 2, \ldots, n$ **do**
      $G_{ij} \leftarrow G_{ij} -$ rowMin$[i]$
  for $i = 1, 2, \ldots, n$ **do**
    for $j = 1, 2, \ldots, n$ **do**
      $G_{ij} \leftarrow G_{ij} -$ colMin$[i]$
  Cover all 0 with the minimum number of horizontal and vertical lines
  $Num \leftarrow$ Minimum number of lines to cover the 0s
Obtain the re-association $\mathbf{G}^*$ and assign the status to users $\rho \leftarrow 1$
for $i = 1, 2, \ldots, |\mathcal{N}|$ **do**
  if $|\mathbf{c}_i| = k_{max}$ **then**
    $\mathcal{N} \leftarrow \mathcal{N} \setminus \mathbf{c}_i$
for $j = 1, 2, \ldots, |\mathcal{M}|$ **do**
  if $\rho = 0$ **then**
    $\mathcal{M}^* \leftarrow \mathcal{M}^* \bigcup m_j$
Add new clusters for UEs in $\mathcal{M}^*$ using Algorithm 1 from Line (4–13)

Fig. 3 describes, in three mobility instances, how the algorithms are handling user dissociation and re-association. In our simulation the mobility of UE is based on random Waypoint model. This movement model applies to non-motorized movements such as walking, running, and cycling. The movement pattern is random both in terms of speed and direction [31]. We initially model a square area in which the BS is at its center. The coverage of the BS is only a fraction of the initial square area forming a circle. As such, the inter cell mobility properties are captured through BS being able to identify the UEs once they enter the coverage area.

5.1 Computational Complexity

In this section, we discuss the computational complexity of the proposed algorithms. Let $g_i^j(n)$ be the computational complexity of the ith algorithm in the jth step. The complexity of Phase 1 of Algorithm 1 can be split as $g_1^1(n) = \mathcal{O}(1)$, $g_1^2(n) = \mathcal{O}(n)$, $g_1^3(n) = \mathcal{O}(n \log n)$, $g_1^4(n) = \mathcal{O}(n \log n)$ and $g_1^5(n) = \mathcal{O}(\log n)$. The complexity in Phase 1 is $\mathcal{O}(n \log n)$. In Phase 2, we get $g_1^1(n) = \mathcal{O}\left(n(n^2 \log n)\right) = \mathcal{O}(n^3 \log n)$, $g_1^2(n) = \mathcal{O}(n)$, $g_1^3(n) = \mathcal{O}(1)$ and $g_1^4(n) = \mathcal{O}(\log n)$. Therefore, the total time complexity for Algorithm 1 is $g_1(n) = \mathcal{O}(n^3 \log n)$. In lines 5-17, Algorithm 2 solves a dynamic constraint satisfaction problem (CSP) through the game theory-based recursive backtracking solution, which solves the problem in polynomial time. For Algorithm 2, the computational complexity is

$$\begin{aligned} g_2(n) &= \mathcal{O}\left(n\left(n\left(n \log n + 1\right) + 1\right)\right) \\ &= \mathcal{O}(n^3 \log n). \end{aligned}$$

In lines 3-23 in Algorithm 3 we use an enhanced brute-force algorithm using heuristics to significantly reduce the search space in each iteration. While we minimize the number of iterations and the search space, the solution is not global. We perform the re-association of users taking each user separately and the computational complexity is

$$\begin{aligned} g_3(n) &= \mathcal{O}\left(n(n(n\log n+1)+1+1\right) \\ &= \mathcal{O}\left(n^3\log n\right). \end{aligned}$$

In lines 18 - 30 of Algorithm 4, the BS solves the Hungarian Algorithm to calculate the clustering order. Algorithm 4 performs the re-association of users using simultaneous mechanism which has the computational complexity:

$$\begin{aligned} g_4(n) &= \mathcal{O}(1+n(n^2\log n+n\log n) \\ &\quad + n(n^2+n^2+1)+n\log n+n\log n) \\ &= \mathcal{O}\left(n^3\log n\right). \end{aligned}$$

5.2 Signalling and Latency

While signaling and processing delay was not the core focus of our study, we believe our methods are efficient and do not add significant latency to the system. 5G is the first generation of mobile networking to use the *inactive* state and it is expected to feature in subsequent generations too. In radio resource control (RRC) *connected* state, the BS saves the scheduling request procedure assuming UE is already synchronized. In RRC *inactive* state, data can be transmitted even without random access channel (RACH) procedure or with 2-step RACH. Therefore, the re-clustering can be implemented in the edge and with limited interaction with the core since the users only change the operating frequency and cluster membership, which can be achieved at the base band unit, closer to the user. The remaining information such as security certificates stay unchanged for the duration of the transmission session. Reducing signalling contributes to power saving and latency reduction.

5.3 Outage probability

We describe the outage probability of the proposed schemes based on the difference between the dissociation and re-association probability as

$$\mathbb{P}(\text{Outage}) = \mathbb{P}(\text{Dissociation}) - \mathbb{P}(\text{Re-association}).$$

The user dissociation probability can be defined as the probability that a user will be dissociated from the cluster due to discrepancies on gain-difference. Suppose that we have n_i users in cluster $\mathbf{c}_i$ and the dissociation probability of the ith user is p_i. Then all users in cluster $\mathbf{c}_i$ will be dissociated with probability given by

$$\begin{aligned} \mathbb{P}(|h_{\mathbf{x}_i} - h_{\mathbf{x}_j}| < \lambda_0^c) &= p_i^{\binom{n_i}{2}} \qquad \forall i,j \in \mathbf{c}_i,\ i \neq j \\ &= p_i^{\left[\frac{n_i(n_i-1)}{2}\right]} \\ &= \exp\left[\frac{n_i(n_i-1)}{2}\ln\left(p_i\right)\right]. \end{aligned} \tag{14}$$

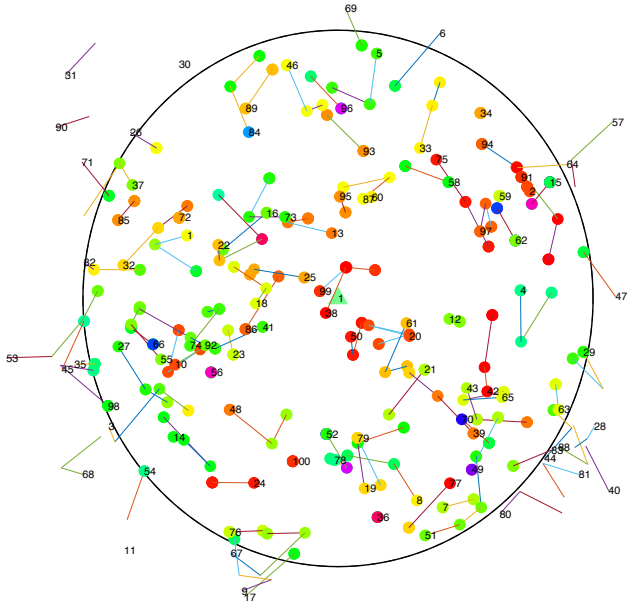

Fig. 3. Various clusters are shown from our simulation. We label each cluster by a distinct color representing its identity. The lines connecting the UEs indicate their movements after three different observations. The change of the color indicates that the node has changed its cluster through dissociation/re-association procedure. The consistent color of moving UEs indicate that the user maintained their original clusters. In this graph, we also show how the users coming in and out of the coverage zone are handled.

For a set of clusters $\{\mathbf{c}_1, \mathbf{c}_2, \ldots, \mathbf{c}_k\} \in \mathcal{C}$, the dissociation probability is the maximum number of users to be dissociated given by

$$\begin{aligned} &\mathbb{P}(\text{all dissociations in } \mathcal{C}) \\ &= \prod_{i=1}^{k} \exp\left[\frac{n_i(n_i-1)}{2}\ln\left(p_i\right)\right] \\ &= \exp\left[\sum_{i=1}^{k}\frac{n_i(n_i-1)}{2}\ln\left(p_i\right)\right]. \end{aligned} \tag{15}$$

The distance between user x_i and x_j $i,j \in \mathbf{c}_i, i \neq j$ is $|h_{x_i} - h_{x_j}|$ is a random variable. The probability p_i is then a function of distance between users, and is defined by

$$p_i = 1 - \exp\left(-\frac{1}{\alpha}\left|h_{x_i} - h_{x_j}\right|\right), \tag{16}$$

where α is the mean of distances between users. Then the dissociation probability in equation (17) becomes

$$\begin{aligned} &\mathbb{P}(\text{all dissociations in } \mathcal{C}) \\ &= \exp\left[\sum_{i=1}^{k}\frac{n_i(n_i-1)}{2}\ \ln\left(1-\exp\left(-\frac{1}{\alpha}\left|h_{x_i}-h_{x_j}\right|\right)\right)\right]. \end{aligned} \tag{17}$$

Now we define the user re-association probability as the probability that a user will be re-associated to a new cluster under the criterion shown in equation (11). Let n_i be the number of users dissociated from cluster $\mathbf{c}_i \in \mathcal{C}, i =$

$1, 2, \ldots, k$. Then the total number of users dissociated is

$$n_1 + n_2 + \cdots + n_k = \sum_{i=1}^{k} n_i = n$$

from clusters $\mathbf{c}_1, \mathbf{c}_2, \ldots, \mathbf{c}_k$, respectively. Through the re-association scheme, the system re-associates all dissociated n users based on the criteria $|h_{\mathbf{x}_i} - h_{\mathbf{x}_j}| > \lambda_1^c$. Let p_i^* be the probability that the ith dissociated user will be re-associated in any of the $\mathbf{c}_j, j = 1, 2, \ldots, k, \forall\ |c_j| < k_{max}$, clusters other than $\mathbf{c}_i$. Let n^* be the number of users satisfying the condition $|h_{\mathbf{x}_i} - h_{\mathbf{x}_j}| > \lambda_1^c$ and $n^* \leq n$. Then, the probability distribution of $n^* = 0, 1, 2, \ldots$ dissociated users re-associated in all the clusters $\mathbf{c}_i, i = 1, 2, \ldots, k$ satisfying the condition follow a binomial distribution and is given by

$$\mathbb{P}(\mathcal{X} = x) = \binom{n}{x} (p_i^*)^x (1 - p_i^*)^{n-x}, \quad x = 0, 1, 2, \ldots, n^*. \tag{18}$$

The re-association probability p_i^* is a random variable and a function of distance between users. The formula of p_i^* can be given by

$$p_i^* = \exp\left(-\frac{1}{\beta} |h_{x_i} - h_{x_j}|\right), \tag{19}$$

where β is the mean of distances between users. Therefore, the re-association probability of n^* users can be written as

$$\mathbb{P}(\mathcal{X} = x) = \binom{n}{x} \left[\exp\left(-\frac{1}{\beta} |h_{x_i} - h_{x_j}|\right)\right]^x \cdot \left[1 - \exp\left(-\frac{1}{\beta} |h_{x_i} - h_{x_j}|\right)\right]^{n-x}, \quad x = 0, 1, 2, \ldots, n^*. \tag{20}$$

However, there is an additional issue to consider, which is the impact of user movement. Let r be the number of users move (leaves or joins) in the cell coverage. The re-association probability for all n dissociated users over the set of clusters $\{\mathbf{c}_1, \mathbf{c}_2, \ldots, \mathbf{c}_k\} \in \mathcal{C}$ is then given by

$$\mathbb{P}(\text{all re-associations in } \mathcal{C} \text{ with } |h_{\mathbf{x}_i} - h_{\mathbf{x}_j}| > \lambda_1^c) = \begin{cases} \binom{n}{x} (p_i^*)^x (1 - p_i^*)^{n-x}, & x = 0, 1, 2, \ldots, (n^* - r) \\ & \text{if } r \text{ users leave} \\ \binom{n}{x} (p_i^*)^x (1 - p_i^*)^{n-x}, & \text{if users remain the same} \\ \binom{n}{x} (p_i^*)^x (1 - p_i^*)^{n-x}, & x = 0, 1, 2, \ldots, (n^* + r) \\ & \text{if } r \text{ users join} \end{cases}$$

where $p_i^* = \exp(-\beta |h_{x_i} - h_{x_j}|)$, n^* is the number of users satisfying the condition $|h_{\mathbf{x}_i} - h_{\mathbf{x}_j}| > \lambda_1^c$ and $n^* \leq n$.

6 MIMO-NOMA

Directional antennas using MIMO can mitigate path loss by focusing the power towards the targeted user. A base station may transmit one or more beam reference signals using Multi-antenna systems. Multi-antenna systems not only provide diversity to individual users but also enables SDMA (space division multiple access), where multiple users can communicate in different beams. MIMO-NOMA overloads SDMA by allocating a cluster of users to each beam and using superposition coding SIC (SC-SIC) within each group [32]. This adds complexity to the clustering problem where users in a cluster not only must conform to the gain-difference but also to the beamwidth and beam direction coverage.

In general, the optimal decoding order for NOMA is the increasing order of the users' channel gains as discussed throughout the paper. However, in the context of NOMA with MIMO beam-forming, the effective channel gains of the users are determined by a combination of the channel gains and the beam-forming directivity gains. We assume that the order of the effective channel gains in a cluster $\boldsymbol{c}$ is

$$|\boldsymbol{h}_1^{\mathrm{H}} \boldsymbol{w}_{\boldsymbol{c}}|^2 > |\boldsymbol{h}_2^{\mathrm{H}} \boldsymbol{w}_{\boldsymbol{c}}|^2 > \ldots > |\boldsymbol{h}_{|\boldsymbol{c}|}^{\mathrm{H}} \boldsymbol{w}_{\boldsymbol{c}}|^2, \tag{21}$$

where $\boldsymbol{h}_1^{\mathrm{H}}$ is the transfer function matrix of the MIMO radio channel, $(\cdot)^{\mathrm{H}}$ is the conjugate transpose and $\boldsymbol{w}_{\boldsymbol{c}} \in \mathbb{C}^N$ is the precoding vector for cluster $\boldsymbol{c}$, which is mapped as a scalar signal to BS N antenna elements. Given the effective MIMO channel gain, the achievable spectral efficiency for cluster c is

$$R_c = B \sum_{i=1}^{m_k} \log_2 \left(1 + \frac{P_i |\boldsymbol{h}_1^{\mathrm{H}} \boldsymbol{w}_{\boldsymbol{c}}|^2}{|\boldsymbol{h}_1^{\mathrm{H}} \boldsymbol{w}_{\boldsymbol{c}}|^2 \sum_{j=1}^{i-1} P_j + \sum_{i \neq c} \sum_{i=1}^{|c|} |\boldsymbol{h}_1^{\mathrm{H}} \boldsymbol{w}_{\boldsymbol{c}}|^2 + \sigma^2}\right) \tag{22}$$

It is also important to clarify that $\boldsymbol{h}$ is derived from the propagation channel by combining the antenna far-field patterns as [33]

$$h_{ij} = \sum_{l=1}^{L} \boldsymbol{h}_{j'}'^{H}(\boldsymbol{\Gamma}_l') \boldsymbol{\alpha}_l \boldsymbol{h}_j(\boldsymbol{\Gamma}_l) \exp(-j2\pi f \tau_l) \tag{23}$$

where $\boldsymbol{\alpha} \in \mathbb{C}^{2\times 2}$ is a polarimetric complex amplitude, $\boldsymbol{\Gamma} = [\phi\ \ \theta]$ and $\boldsymbol{\Gamma}' = [\phi'\ \ \theta']$ are vectors composed of azimuth and polar angles of the plane wave radiation and reception at the transmitter and receiver, respectively, and τ is the propagation delay time. $\boldsymbol{h}(\boldsymbol{\Gamma}) \in \mathbb{C}^2$ is polarimetric complex gain of the antenna element to the direction $\boldsymbol{\Gamma}$. The phase of the antenna elements are defined with respect to an origin of the coordinate system of the radio channel. The polar angle θ is related to the elevation angle ψ as $\theta = \pi/2 - \psi$. A symbol with subscript $(\cdot)_l$ means a parameter value for the lth plane wave, while the quantity represents the transmitter side if a symbol is with a prime $(\cdot)'$.

While serving users through MIMO beamforming offers high link strength through compensating for path loss and suppressing co-channel interference, it is less resilient to highly mobile users, such that, a slight misalignment with the serving antenna array may completely break the link, causing severe outages. Misalignment results from radial mobility of the users around the transmitter. The problem is exacerbated in MIMO-NOMA, where users in a cluster are served by a single narrow beam. In addition, although the signal strength is maximum at perfect alignment – at the center of the beam – it decreases steeply as the

TABLE 2
Simulation parameters and their values

Parameter	Value
Network area	500 × 500 meters squared
Cell radius	200 meters
Users in the network	600
Max Transmit Power	46 dBm
User receive antenna gain	0 dBi
BS transmit antenna gain	0 dBi
AWGN	-90 dBm
Bandwidth (B)	1 Hz
Mobility intensity	[20 - 80] %

receiver moves along the circumference of the BS, even when still within the azimuth of the antenna beam, relative to the beamwidth of the array elements. To address these problems effectively, we consider NOMA system with adaptive discrete beam-switching, assisted with the KMAA algorithm. We adopt switched beam technique due to its low complexity and high performance [34]. Beam-steering and beam-adaptation are among the possible ways to improve resilience in MIMO-NOMA, however, they are beyond the scope of this paper. For simplicity, in this work we assume that there is no overlap between beams and we consider a beam model where the gain within the beamwidth is non-uniform and zero outside. To integrate MIMO-NOMA beamforming into our KMAA algorithm, we make several changes which we summarize below:

1) deploy dissociation/re-association mechanism, while clusters are formed to maximize the gain-difference between users in a cluster using the KMAA algorithm.
2) dissociate users that move outside the serving beam borders.
3) dissociate users who move within the beam borders but fall below the gain-difference threshold with one or more users.
4) consider all dissociated users and newly identified users in the cell coverage. Users which are dissociated but fall outside the omni-directional discovery range of a BS are no longer considered in the optimization.
5) the best sector for each user is searched against its quasi-omni patterns.
6) deploy the KMAA algorithm to find the global optimal solution within each beam for each sector.
7) assign a new cluster with different beam and channel to each dissociated user, if one is available and has spare capacity.
8) if there are no active beams available and/or existing beams have no spare capacity, then a new beam is formed, assuming there are enough antennas.

7 Simulation Results

The cluster size has impact on the number of dissociation and re-association. The density of users in the network increases or decreases the number of dissociations and re-associations, which in turn, impact the latency perceived by the users. This latency is made up largely of the transition duration between dissociation and re-association of a user. The increase in search time for a new cluster will certainly impact the outages. In addition, the percentage of mobile users in the network will also influence the dissociation and re-association rate in the system. Our goals in this section is to explore the effectiveness of the proposed clustering algorithms in minimizing the number of clusters and the number of re-association by implementing the algorithms and the optimizations in our simulation. We also analyze various other metrics including throughput and outage probability.

Experiment setup: In the simulation, we consider an area of 500 × 500 meters squared with a total of 800 users, randomly dropped with uniform distribution. The area is covered by one BS, located at the center, with coverage radius of 200 meters. For these experiments we consider a fraction of the users to be mobile, using Random Waypoint mobility model. The mobile users can randomly move in or out of the BS coverage, but remain inside the defined area. This is to test our solutions in handling inter and intra-cell mobility. Further simulation parameters are given in Table 7.

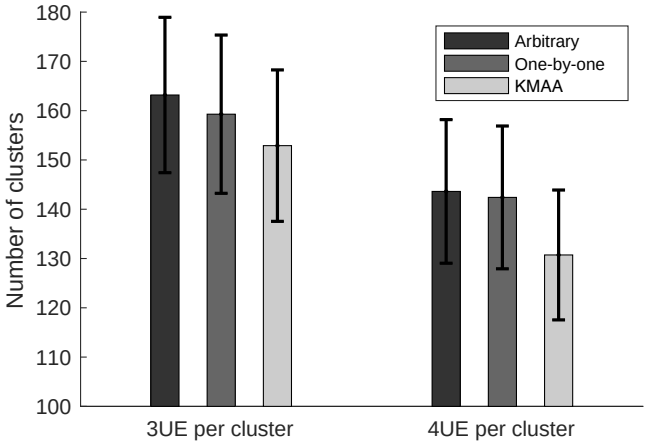

Fig. 4. Number of clusters with maximum cluster size of 3 and 4.

Channel gain: There are many path-loss models available in the literature obtained through real-world measurements in urban, suburban and rural areas, based on probabilistic approaches, designed for different frequencies and environments [35], [36], [37]. To predict the channel gain between the BS and users, we used the ABG model for Urban Micro (UMi) under Non-Line-of-Sight (NLOS) that describes large-scale propagation path loss at sub-6GHz and millimeter-wave frequencies [38], [39].

$$\begin{aligned} \mathrm{PL[dB]} = 10\,\alpha\,\log_{10}\left(\frac{d_i}{1\mathrm{m}}\right) + \beta \\ + 10\,\gamma\,\log_{10}\left(\frac{f}{1\mathrm{GHz}}\right) + \chi_{\sigma}^{\mathrm{ABG}}, \end{aligned} \tag{24}$$

where α and γ are coefficients showing the dependence of path loss on distance and frequency, respectively, β is an optimized offset value for path loss in dB, f is the carrier frequency in GHz, and $\chi_{\sigma}^{\mathrm{ABG}}$ is the shadow fading

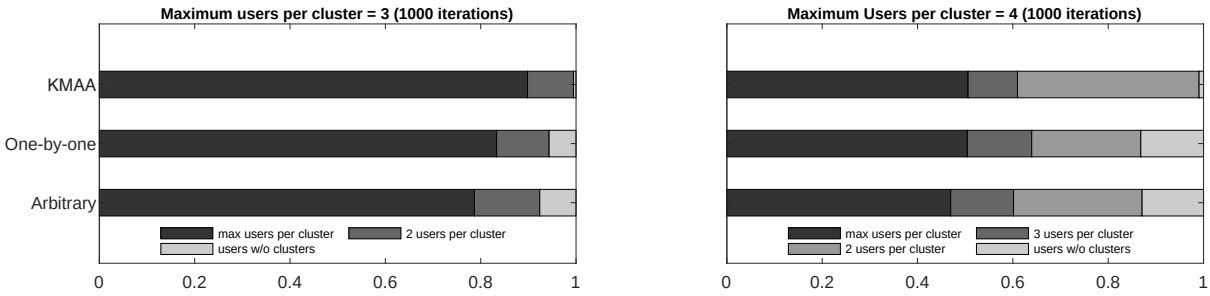

Fig. 5. Percentage of number of users in clusters using arbitrary, one-by-one and the KMAA algorithm for (left) Maximum users per cluster = 3 and (right) Maximum users per cluster = 4.

(SF) standard deviation, describing large-scale signal fluctuations about the mean path loss over distance. The following parameters are used in the simulation: α = 3.5, β = 24.4 dB, γ = 1.9, σ = 8 dB and f = 3.5 GHz. The total transmission power of the BS is 46 dBm. The antenna gain at the BS and UE is 0 dBi. A one-antenna transmission and one-antenna reception (1-by-1 SISO) system is assumed. We have built a simulation flexible enough to evaluate our solutions against network variations, enabling the configuration of cluster size, number of users in the cell, percentage of the mobile users and size of the cell.

7.1 Impact of mobility on the number of clusters

Due to the fact that the proposed solutions manage clusters differently, the number of clusters produced by each algorithm varies. The number of clusters is directly proportional to the number of required resource blocks. In Fig. 4 we provide analysis for each method based on 1000 mobility instances, given two cluster sizes: 3 and 4. As can be seen in the Fig., with this metric, and with 443 average number of users under the coverage area, the number of clusters formed by the KMAA algorithm is considerably lower. This can be explained by the fact that the KMAA algorithm maintains optimal gain-difference across all clusters simultaneously, while the other two algorithms are less optimal in this respect.

7.2 Impact of the proposed algorithms on clustering efficiency

Fig. (5-left) compares the ratio of number of users per cluster of the three proposed algorithms, given a maximum cluster size of 3. The arbitrary algorithms is the simplest of the three algorithms, but results in a degraded performance, where more than 20% of the clusters are underutilized, compared to under 10% for the KMAA. When the maximum number of users per cluster is set to 4, the performance of the arbitrary is even worse with more than half the clusters not reaching full capacity, as shown in Fig. (5-right). On the other hand, the one-by-one and the KMAA are comparable, with a slight advantage in favor of the latter. The KMAA clustering algorithm represent the most efficient solution in handling mobility, achieving an improved clustering efficiency, over the heuristic-based algorithms which can not provide approximation guarantee. The clustering efficiency translates to increased number of served users, given a fixed set number of frequency blocks.

Clustering efficiency has an impact on CPU cost too. We have computed the processing time of each method under different mobility intensities (40% and 80%) and maximum cluster sizes of 3 and 4. The measurements were performed using MATLAB 2020b running on a MacBook Pro with 2.2 GHz 6-Core Intel Core i7 processor and 16 GB of memory (see Table 3). The processing time includes collating measurements from UEs, running algorithms and reclustering. We show that decreasing mobility from 80% to 40% consistently reduces CPU cost when the KMAA algorithm is used, while it is not always the case when using Arbitrary and One-by-one algorithms, highlighting another advantage of our proposed algorithm. The lack of a global optimal solution in the Arbitrary and One-by-one algorithms reduces clustering accuracy and leads to increased CPU cost. We also note that the transition time of moving one user from one cluster to another is only the cost of transmitting and processing a few bits, which does not incur any meaningful overhead or delay to the network.

In our proposed methods, the BS only leverages the channel quality indicator (CQI), which the UEs transmit periodically or aperiodically through control channels to feedback channel state information (CSI) every few milliseconds. Based on these statistics the BS continuously determines the clusters order and performs dissociations/association procedures. As such, our method does not add significant overhead in terms of signalling with the BS or core network.

7.3 Impact of the proposed algorithms on the number of re-associations

In large and crowded scenarios with mobility, the number of decisions the BS is required to make to find alternative position is time consuming and leads to higher latency which deteriorates the overall system performance. It can trigger multiple re-transmission timeouts and a slow growth of the congestion window. In this section, we analyze the efficiency of our methods in respect to the number of required re-association. In Fig. 6, we analyze the number of re-associations, given two maximum cluster sizes: 3 and 4, by plotting the cumulative probability against the number of re-associations. The re-association rate affects different system performance metrics such as signaling load and user perceived quality of experience (QoE). When the cluster size is 3 (Fig. 6-left) the KMAA and the one-by-one algorithms are competitive and both methods yield better results than the arbitrary. KMAA can reduce disscociation/association

TABLE 3
Averaged CPU cost per mobility instance with 400 users in the network.

	Arbitrary	One-by-one	KMAA	Arbitrary	One-by-one	KMAA
Mobility intensity			80%			
Maximum cluster size		3			4	
CPU time per iteration in seconds	0.2862	0.2738	0.2964	0.364	0.4002	0.3968
Mobility intensity			40%			
Maximum cluster size		3			4	
CPU time per iteration in seconds	0.2766	0.3112	0.2448	0.39	0.431	0.351

events by approximately 28.5% when compared to the arbitrary method. By comparison, the KMAA algorithm is confirmed as the best variant of the three algorithms when we set the maximum cluster size to 4. The low number of re-associations implies shorter transition duration time between dissociation and re-association of a user.

7.4 Clustering comparison

Jaccard coefficient is a commonly applied statistical indicator for measuring the pairwise similarity [40]. We leverage the method to compare the differences in clustering of the three solutions. For two cluster sets $\mathcal{C}_i$ and $\mathcal{C}_i^*$, it can be defined as the ratio of the number of elements in their intersection against the number of elements in their union, and is given by

$$\begin{aligned} Q &= \frac{1}{|\mathcal{C}|}\sum_{i=1}^{|\mathcal{C}|} J(\mathcal{C}_i, \mathcal{C}_i^*) \\ &= \frac{1}{|\mathcal{C}|}\sum_{i=1}^{|\mathcal{C}|} \frac{|\mathcal{C}_i \bigcap \mathcal{C}_i^*|}{|\mathcal{C}_i \bigcup \mathcal{C}_i^*|} \end{aligned} \tag{25}$$

where $J(\mathcal{C}_i, \mathcal{C}_i^*) \in [0, 1]$ is the Jaccard coefficient and $|\cdot|$ denotes the cardinality of a set. The Jaccard index has a value of 0 when the two clusters have no UEs in common, 1 when they have exactly the same UEs, and strictly between 0 and 1 otherwise. The two sets are more similar (have more common elements) when the value gets closer to 1. Fig. 7 shows the average Jaccards similarity per cluster between the three clustering methods under 25 mobility instances. From each clustering method we pull all the clusters and make cross-comparisons. From the Fig. it become clear that the similarity between clustering diminishes as users continue to move. It also shows that the clustering differences between the KMAA and arbitrary is the greatest. This clearly implies that the achievable capacity of each method is varied. We study the impact of different clustering techniques in the following section.

7.5 Impact on throughput

In general, the clustering algorithms presented in this paper are used to determine which users should be assigned to which cluster, such that the SIC is always achievable. However, the optimization in $\mathcal{P}_2$ is oblivious to the throughput of clusters. In practice, different clusters may yield significantly different data rate gains. The variations in clustering outcome of each method has impact on the overall throughput of the network. This is because the gain-difference between users can vary from cluster to another and more prominent in highly dynamic networks. In this section we analyze our clustering methods in respect to achievable capacity in bits/sec/hertz. We use the following formula to evaluate the rate per cluster [41]:

$$R_c = B\sum_{i=1}^{m_k}\log_2\left(1 + \frac{P_i\,|h_i|^2}{\sum\limits_{j=1}^{i-1} P_j\,|h_i|^2 + \sigma^2}\right) \tag{26}$$

where $|h_i|^2$ is the normalized channel gain of the ith user. The additive white Gaussian noise (AWGN) is assumed to be normalized with zero mean and variance σ^2 and B is the size of the resource block. $\sum_{j=1}^{i-1} P_j|h_i|^2$ is the inter-user interference for ith user in downlink cluster. The total capacity of the cell is given by

$$R = \sum_{k=1}^{|\mathcal{C}|}\left(B\sum_{i=1}^{m_k}\log_2\left(1 + \frac{P_i\,|h_i|^2}{\sum\limits_{j=1}^{i-1} P_j\,|h_i|^2 + \sigma^2}\right)\right). \tag{27}$$

Note that, if the NOMA user can not be assigned to a suitable cluster, we assume the traditional time division multiple access (TDMA) scheme where

$$\sum_{j=1}^{0} P_j\,|h_i|^2 = 0. \tag{28}$$

TABLE 4
Capacity gain of the KMAA over arbitrary and one-by-one.

Method	Threshold (dissociation, re-association)					
	(2dB, 3dB)			(4dB, 5dB)		
	3UE	4UE	5UE	3UE	4UE	5UE
KMAA vs Arbitrary	1.06x	1.21x	1.29x	1.41x	1.61x	1.68x
KMAA vs One-by-one	1.09x	1.24x	1.33x	1.45x	1.63x	1.66x

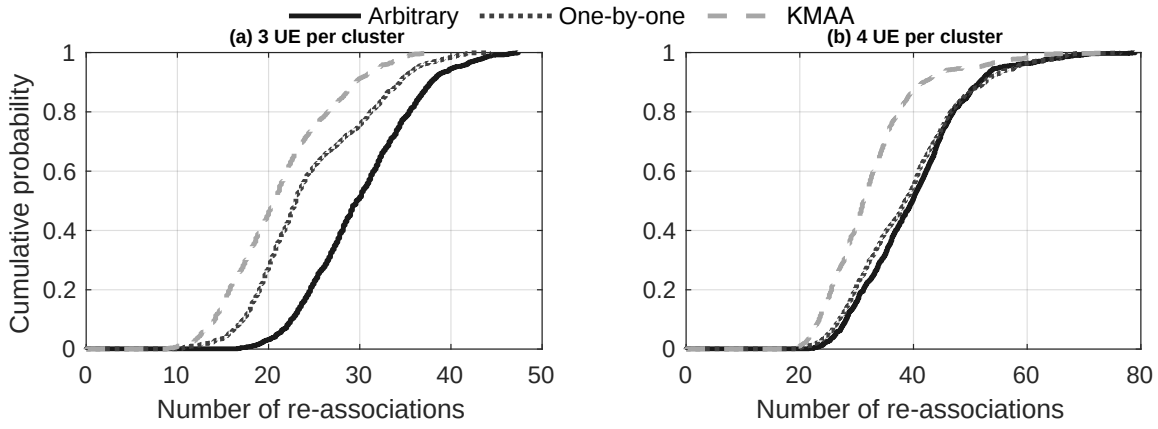

Fig. 6. Simulation results for comparing the KMAA algorithm with arbitrary and one-by-one in terms of number of re-associations. (left) 3 UEs per cluster and (right) 4 UEs per cluster. The KMAA algorithm reduces the number of re-associations when cluster size is higher than 3.

The corresponding receive rates in the TDMA system is computed as

$$R_i = B \log_2 \left(1 + \frac{P_i |h_i|^2}{\sigma^2}\right). \quad (29)$$

Fig. 8 compares the capacity of the arbitrary, one-by-one and the KMAA. The network contains 600 users and we simulated 50 mobility instances. The threshold for dissociation and re-association is 2dB and 3dB for top row (4dB and 5dB (bottom row)), respectively. Achievable throughput of arbitrary is only a fraction less than the throughput of one-by-one algorithm. We observe that the total capacity of the KMAA algorithm is much higher than the total capacity of the other two algorithms when the maximum cluster size is set to 4 and 5. Table 4 summarizes the gain of the KMAA over the arbitrary and one-by-one, showing that the latter is the preferred clustering method choice for NOMA system design.

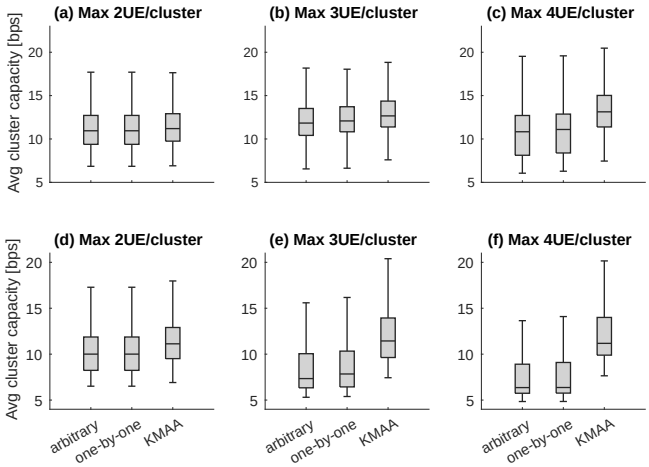

Fig. 8. Comparative average cluster capacity performance with maximum 3, 4 and 5 users per cluster with (top row) λ_0 = 2dB and λ_1 = 3dB and (bottom row) λ_0 = 4dB and λ_1 = 5dB.

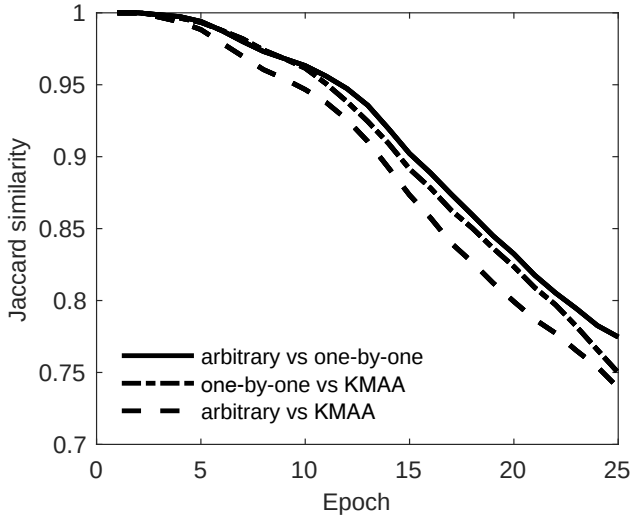

Fig. 7. Comparison of the Jaccard-based similarity of clusters for arbitrary, one-by-one and the KMAA algorithms.

7.6 Varying the degree of mobility

If a continuous channel feedback is available at the base station, UEs are clustered according to their instantaneous channel conditions. Under average channel condition information with inefficient clustering, which may result in cases where the channel gain information is not instantly available in line with the scheduled periodic reporting mechanism, typically every 2ms, the outage probability increases with the mobility of users [42]. In this section, we investigate the average number of dissociations and re-associations incurred by using each of the proposed methods. To evaluate our methods more rigorously and to understand how our proposed algorithms work in a wide range of mobility scenarios we vary the degree of mobility in our simulation from 20% to 80%. In this experiment, a fraction of users are set to be mobile for a consecutive series of 150 movements in various directions. The speeds are set between 1m and 10m per movement. It can be easily observed that in all cases considered, our methods work well under different wireless environments and different levels of mobilities as shown in Figure 9. At the low end, we notice the dissociation and

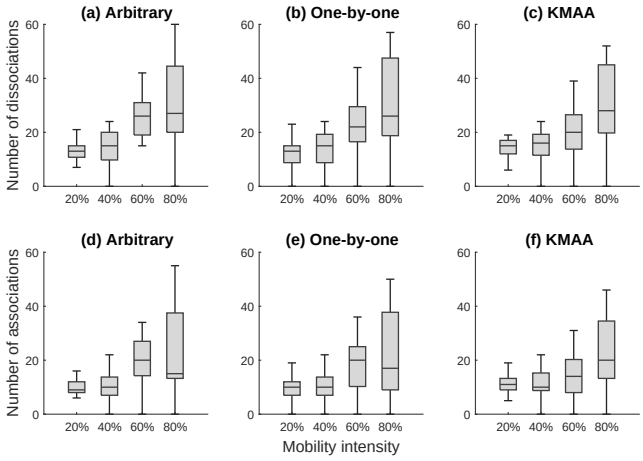

Fig. 9. Comparative number of dissociations (top row) and re-associations (bottom row) performance between arbitrary, one-by-one and the KMAA under various mobility intensity with 600 users and maximum 3 users per cluster.

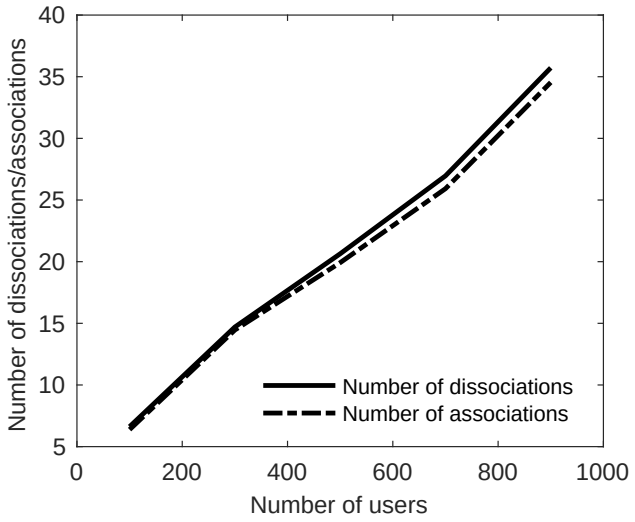

Fig. 10. Number of dissociations and re-associations using the KMAA under various number of users with 80% mobility intensity and maximum 3 users per cluster.

re-association rate is lowest under all solutions. This is due to the fact that the gain-difference between users remain adequately large to permit SIC. As the intensity of mobility increase, we see higher dissociation and re-association rate, but noticeably, the KMAA algorithm shows the lowest rate, indicating superior clustering efficiency and more resilience to delay in feedback compared to other methods.

We have also analyzed the impact of different numbers of users within the network with 80% mobility intensity and a maximum of 3 users per cluster for the KMAA algorithm. It is observed that the number of dissociations and associations are directly proportional to the number of users in the network, as seen in Figure 10.

7.7 Impact of mobility intensity on outage probability

In this section, the outage probability, which corresponds to the three methods is evaluated through extensive simulations over 150 mobility instances. Table 5 presents the outage probabilities for various mobility intensities between 20% to 80%, considering maximum 3 UE per cluster. Our results illustrate that the outage performance for the KMAA algorithms outperforms the other two solutions by at least 89% when the mobility intensity is set to 80%. We also note that the outage probability is significantly lower (ranging between 0.28% and 0.38% for the KMAA method) compared to outage probability resulting from other factors such as decoding, transmission, and collision, highlighting suitability for practical implementation [43], [44], [45].

TABLE 5
Outage probabilities under various mobility scenarios.

Method	Mobility Intensity			
	20%	40%	60%	80%
Arbitrary	0.0351	0.0353	0.0322	0.0357
One-by-one	0.0307	0.0291	0.0272	0.0307
KMAA	0.0028	0.0031	0.0036	0.0038

Implications: From our analysis we can conclude that the number of clusters can be lowered by using the KMAA algorithm, especially in the case where the cluster sizes is greater than 3. Nevertheless, our analysis can also provide indications on selecting the appropriate maximum cluster size, given the size of the network and other parameters. Through these results, we have shown that these techniques can be used in NOMA clusters to manage mobility instead of complex power allocation techniques, such as cognitive radio power allocation, with high computational costs, which has frequently been cited as a major shortcoming of SIC [46], [47].

The clusters arrangement vary according to the frequency band used for transmissions, since the channel to and from the user change with the frequency. In cellular networks, the uplink and the downlink operate on different frequencies (over 20 MHz apart). One point that is not particularly addressed in this paper is uplink NOMA. However, the clustering techniques presented in this paper is not restricted to downlink NOMA only, the same methods could also be applied to uplink NOMA transmissions, with minor changes to the algorithms.

7.8 Link resilience of the KMAA algorithm under NOMA-MIMO beam-switching

We compare the link resilience of the KMAA algorithm under beam-switching considering 3 different beamwidths: $45°$, $22.5°$ and $11.25°$ as in [48]. Fig. 11 shows the cumulative probability of switching rate for (left) 40% mobility and (right) 80% mobility. In both figures (left and right), we can see that narrower beams exhibit more beam switching, under constant speed. The level of mobility also has an impact on the switching rate, where nearly 70% of the UEs perform beam switching after 150 mobility instances when 80% of UE are mobile compared to just under 35% with the level of mobility set to 40%. The analysis shows that there is a fundamental trade-off in the design of beamwidth in NOMA-MIMO in mobility scenarios: wider beams may suffer from

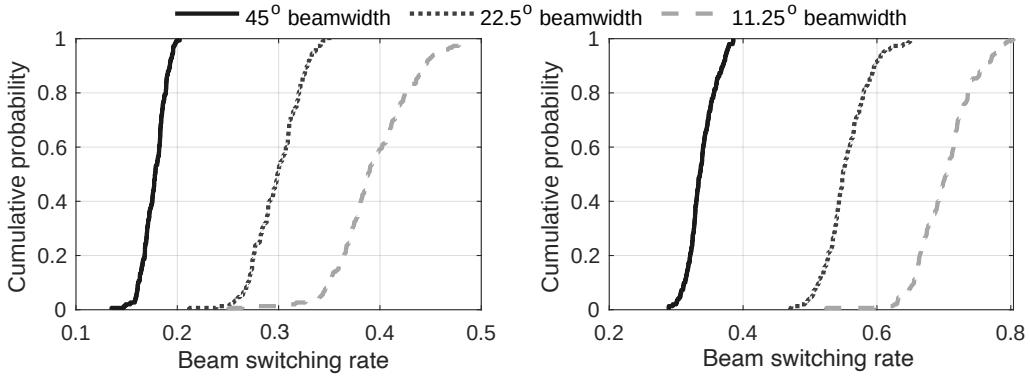

Fig. 11. Simulation results for comparing the KMAA algorithm with. (left) 40% mobility rate and (right) 80% mobility rate.

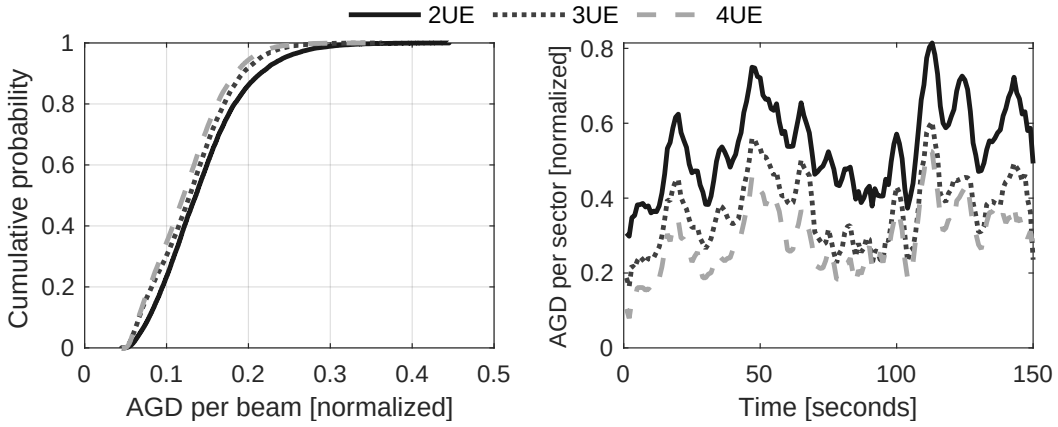

Fig. 12. Simulation results for comparing the KMAA algorithm with Average Gain Difference (AGD) 80% mobility.

inefficient spectrum utilization and lower throughput but narrower beams are less resilient to mobility,and may add latency and overhead to the system.

7.9 Average gain-difference between users in clusters

The KMAA algorithm with mobility support aims to maximize channel gain-difference while minimizing NOMA interference. We analyzed the Average Gain Difference (AGD) under 80% mobility with 2, 3 and 4 users per cluster. We first looked at the cumulative probability under 150 mobility instances. From Fig. (11 (left)) we observe that the average gain difference is maximum when the maximum number of users per cluster is set to 2. However the difference is not significant when the maximum cluster size is set to 3 and 4. This is further reinforced when we evaluate the AGD per beam as shown in Fig. (11 (right)).

7.10 Discussion

Power allocation: Our proposed clustering solutions are performed solely using dissociation/association procedure, as described in the algorithms. The BS allocates a fixed amount of transmit power to each user (aka fixed power NOMA) [7]. Based on our results, the proposed techniques achieve good performance without the need for power allocation optimization. However, power allocation could also be incorporated to our management system to further improve the performance which in turn can reduce the number of dissociation events as well as the number of beam-switching and handovers.

Cluster size: It is to be noted that larger cluster size (> 3) requires more computational power and has impact on bit error rate to cater for longer decoding delay as the SIC latency linearly increases with the cluster size. However, recent studies have shown promising results for larger cluster sizes (up to 6 users per cluster) [18]. While we indicate to 3 and 4 cluster size in the results, this is only when optimization conditions are satisfied (as per algorithm 2, 3 and 4). And as shown in Fig. 5 (right), only around 50 percent of the total clusters are reaching the maximum allowed cluster size, 4. This can be further controlled by setting larger gain-difference threshold. Additionally, an emerging topic in the literature where broadcast scenario involving IoT devices is proving to be a strong proposition for having several users in a cluster where IoT devices do not perform SIC in the enhance layer [49]. We believe such new emerging applications and promising results are indicators to be considered.

8 Conclusion

In this work we highlighted the problem of mobility associated with NOMA clustering, where the gain-difference between users may decrease to a level, where the successive interference cancellation (SIC) fails. This in turn diminishes the NOMA performance gain over OMA. In this context we presented a new approach, fully-automatic, to manage and update clusters in a robust manner, through user dissociation and re-association procedure, which links dissociated users to new clusters. We have analyzed three solutions: (i) arbitrarily (ii) one-by-one and (iii) KMAA. We conducted extensive experiments through simulations to compare the performances of each solution. We showed that the solutions efficiently maintain the clusters with the desired gain-difference among users to facilitate SIC decoding. We note here that the proposed KMAA algorithm remarkably improves outage probability and global throughput when deployed in dense urban and highly mobile environments, without increasing time complexity. We further showed that the Hungarian-powered solution can provide an enhanced link resiliency in the face of mobility under various beam widths. We believe that this clustering solution is the first in this field. It is also generic and can be extended to cover several other kinds of networks underpinned by NOMA technology, where mobility handling is essential, e.g., connectivity of autonomous and semi-autonomous vehicles. In the future, we aim to investigate the impact of the proposed clustering techniques considering inter-cell user mobility. Another interesting direction is to find the optimal threshold value for dissociation and association.

References

[1] N.-S. Vo, T. Q. Duong, M. Guizani, and A. Kortun, "5G optimized caching and downlink resource sharing for smart cities," *IEEE Access*, vol. 6, pp. 31 457–31 468, 2018.

[2] M. Erol-Kantarci and S. Sukhmani, "Caching and computing at the edge for mobile augmented reality and virtual reality (AR/VR) in 5G," in *Ad Hoc Networks*. Springer, 2018, pp. 169–177.

[3] J. M. Meredith, "3gpp tr 38.812 v16.0.0- 3rd generation partnership project; technical specification group radio access network; study on non-orthogonal multiple access (noma) for nr (release 16)," *Tech. Rep.*, 2018.

[4] Y. Saito, Y. Kishiyama, A. Benjebbour, T. Nakamura, A. Li, and K. Higuchi, "Non-orthogonal multiple access (NOMA) for cellular future radio access," in *Vehicular Technology Conference (VTC Spring), 77th*. IEEE, 2013, pp. 1–5.

[5] R. Abozariba, M. K. Naeem, M. Patwary, M. Seyedebrahimi, P. Bull, and A. Aneiba, "NOMA-based resource allocation and mobility enhancement framework for IoT in next generation cellular networks," *IEEE Access*, vol. 7, pp. 29 158–29 172, 2019.

[6] J. Choi, "On the power allocation for a practical multiuser superposition scheme in NOMA systems," *IEEE Communications Letters*, vol. 20, no. 3, pp. 438–441, 2016.

[7] Z. Ding, P. Fan, and H. V. Poor, "Impact of user pairing on 5G non-orthogonal multiple-access downlink transmissions," *IEEE Transactions on Vehicular Technology*, vol. 65, no. 8, pp. 6010–6023, 2015.

[8] Z. Ding, M. Peng, and H. V. Poor, "Cooperative non-orthogonal multiple access in 5G systems," *IEEE Communications Letters*, vol. 19, no. 8, pp. 1462–1465, 2015.

[9] B. Kimy, S. Lim, H. Kim, S. Suh, J. Kwun, S. Choi, C. Lee, S. Lee, and D. Hong, "Non-orthogonal multiple access in a downlink multiuser beamforming system," in *IEEE Military Communications Conference (MILCOM)*. IEEE, 2013, pp. 1278–1283.

[10] H. Al-Zubaidy, J. Liebeherr, and A. Burchard, "Network-layer performance analysis of multihop fading channels," *IEEE/ACM Transactions on Networking*, vol. 24, no. 1, pp. 204–217, 2014.

[11] M. J. Neely, "Dynamic power allocation and routing for satellite and wireless networks with time varying channels," Ph.D. dissertation, Massachusetts Institute of Technology, 2003.

[12] H.-R. Kim, J. Chen, and J. Yoon, "Joint User Clustering and Beamforming in Non-Orthogonal Multiple Access Networks," *IEEE Access*, vol. 8, pp. 111 355–111 367, 2020.

[13] M. K. Haider and E. W. Knightly, "Mobility resilience and overhead constrained adaptation in directional 60 GHz WLANs: protocol design and system implementation," in *Proceedings of the 17th ACM International Symposium on Mobile Ad Hoc Networking and Computing*, 2016, pp. 61–70.

[14] M. K. Naeem, R. Abozariba, M. Asaduzzaman, and M. Patwary, "Towards the mobility issues of 5g-noma through user dissociation and re-association control," in *2020 IEEE 21st International Symposium on" A World of Wireless, Mobile and Multimedia Networks"(WoWMoM)*. IEEE, 2020, pp. 427–432.

[15] Z. Ding, L. Dai, and H. V. Poor, "Mimo-noma design for small packet transmission in the internet of things," *IEEE access*, vol. 4, pp. 1393–1405, 2016.

[16] S. R. Islam, M. Zeng, O. A. Dobre, and K.-S. Kwak, "Resource allocation for downlink NOMA systems: Key techniques and open issues," *IEEE Wireless Communications*, vol. 25, no. 2, pp. 40–47, 2018.

[17] M. Diamanti, G. Fragkos, E. E. Tsiropoulou, and S. Papavassiliou, "Unified user association and contract-theoretic resource orchestration in NOMA heterogeneous wireless networks," *IEEE Open Journal of the Communications Society*, vol. 1, pp. 1485–1502, 2020.

[18] A. Celik, M.-C. Tsai, R. M. Radaydeh, F. S. Al-Qahtani, and M.-S. Alouini, "Distributed user clustering and resource allocation for imperfect NOMA in heterogeneous networks," *IEEE Transactions on Communications*, vol. 67, no. 10, pp. 7211–7227, 2019.

[19] ——, "Distributed cluster formation and power-bandwidth allocation for imperfect NOMA in DL-HetNets," *IEEE Transactions on Communications*, vol. 67, no. 2, pp. 1677–1692, 2018.

[20] J. Cui, Z. Ding, P. Fan, and N. Al-Dhahir, "Unsupervised machine learning-based user clustering in millimeter-wave-NOMA systems," *IEEE Transactions on Wireless Communications*, vol. 17, no. 11, pp. 7425–7440, 2018.

[21] S. Ali, E. Hossain, and D. I. Kim, "Non-orthogonal multiple access (NOMA) for downlink multiuser MIMO systems: User clustering, beamforming, and power allocation," *IEEE access*, vol. 5, pp. 565–577, 2016.

[22] Z. Chen, Z. Ding, and X. Dai, "Beamforming for combating inter-cluster and intra-cluster interference in hybrid NOMA systems," *IEEE Access*, vol. 4, pp. 4452–4463, 2016.

[23] H. Zhang, H. Zhang, W. Liu, K. Long, J. Dong, and V. C. Leung, "Energy efficient user clustering, hybrid precoding and power optimization in terahertz MIMO-NOMA systems," *IEEE Journal on Selected Areas in Communications*, vol. 38, no. 9, pp. 2074–2085, 2020.

[24] E. Hyytiä and J. Virtamo, "Random waypoint mobility model in cellular networks," *Wireless Networks*, vol. 13, no. 2, pp. 177–188, 2007.

[25] D. Tse and P. Viswanath, *Fundamentals of wireless communication*. Cambridge university press, 2005.

[26] E. A. Sourour and M. Nakagawa, "Performance of orthogonal multicarrier cdma in a multipath fading channel," *IEEE transactions on communications*, vol. 44, no. 3, pp. 356–367, 1996.

[27] R. Matsushita and T. Tanaka, "Low-rank matrix reconstruction and clustering via approximate message passing," in *Advances in Neural Information Processing Systems*, 2013, pp. 917–925.

[28] H. W. Kuhn, "The Hungarian method for the assignment problem," *Naval research logistics quarterly*, vol. 2, no. 1-2, pp. 83–97, 1955.

[29] "3GPP Release 16: Study items and road map," http://download.ni.com/evaluation/rf/33656_3GPP_Release_16_WP_Ltr_WR.pdf, 2018.

[30] A. Kiani and N. Ansari, "Edge computing aware NOMA for 5G networks," *IEEE Internet of Things Journal*, vol. 5, no. 2, pp. 1299–1306, 2018.

[31] A. K. Maurya, D. Singh, A. Kumar, and R. Maurya, "Random Waypoint mobility model based performance estimation of On-Demand routing protocols in MANET for CBR applications," in *2014 International Conference on Computing for Sustainable Global Development (INDIACom)*. IEEE, 2014, pp. 835–839.

[32] M. Vaezi, R. Schober, Z. Ding, and H. V. Poor, "Non-orthogonal multiple access: Common myths and critical questions," *IEEE Wireless Communications*, vol. 26, no. 5, pp. 174–180, 2019.

[33] K. Haneda, "Channel models and beamforming at millimeter-wave frequency bands," *IEICE Transactions on Communications*, vol. 98, no. 5, pp. 755–772, 2015.

[34] J. Kim and A. F. Molisch, "Enabling Gigabit services for IEEE 802.11 ad-capable high-speed train networks," in *2013 IEEE Radio and Wireless Symposium*. IEEE, 2013, pp. 145–147.

[35] G. R. MacCartney and T. S. Rappaport, "Rural macrocell path loss models for millimeter wave wireless communications," *IEEE Journal on Selected Areas in Communications*, vol. 35, no. 7, pp. 1663–1677, 2017.

[36] S. Y. Seidel and T. S. Rappaport, "914 MHz path loss prediction models for indoor wireless communications in multifloored buildings," *IEEE transactions on Antennas and Propagation*, vol. 40, no. 2, pp. 207–217, 1992.

[37] J. Wu, S. Rangan, and H. Zhang, *Green communications: theoretical fundamentals, algorithms, and applications*. CRC press, 2016.

[38] K. Haneda, J. Zhang, L. Tan, G. Liu, Y. Zheng, H. Asplund, J. Li, Y. Wang, D. Steer, C. Li *et al.*, "5G 3GPP-like channel models for outdoor urban microcellular and macrocellular environments," in *IEEE 83rd Vehicular Technology Conference (VTC Spring)*. IEEE, 2016, pp. 1–7.

[39] S. Sun, T. S. Rappaport, S. Rangan, T. A. Thomas, A. Ghosh, I. Z. Kovacs, I. Rodriguez, O. Koymen, A. Partyka, and J. Jarvelainen, "Propagation path loss models for 5G urban micro-and macro-cellular scenarios," in *IEEE 83rd Vehicular Technology Conference (VTC Spring)*. IEEE, 2016, pp. 1–6.

[40] P. Jaccard, "The distribution of the flora in the alpine zone. 1," *New phytologist*, vol. 11, no. 2, pp. 37–50, 1912.

[41] M. S. Ali, H. Tabassum, and E. Hossain, "Dynamic user clustering and power allocation for uplink and downlink non-orthogonal multiple access (NOMA) systems," *IEEE access*, vol. 4, pp. 6325–6343, 2016.

[42] S. Timotheou and I. Krikidis, "Fairness for non-orthogonal multiple access in 5G systems," *IEEE Signal Processing Letters*, vol. 22, no. 10, pp. 1647–1651, 2015.

[43] Z. Ding, Z. Yang, P. Fan, and H. V. Poor, "On the performance of non-orthogonal multiple access in 5g systems with randomly deployed users," *IEEE signal processing letters*, vol. 21, no. 12, pp. 1501–1505, 2014.

[44] H. Sun, Q. Wang, R. Q. Hu, and Y. Qian, "Outage probability study in a noma relay system," in *2017 IEEE Wireless Communications and Networking Conference (WCNC)*. IEEE, 2017, pp. 1–6.

[45] X. Liang, Y. Wu, D. W. K. Ng, Y. Zuo, S. Jin, and H. Zhu, "Outage performance for cooperative noma transmission with an af relay," *IEEE Communications Letters*, vol. 21, no. 11, pp. 2428–2431, 2017.

[46] S. R. Islam, N. Avazov, O. A. Dobre, and K.-S. Kwak, "Power-domain non-orthogonal multiple access (NOMA) in 5G systems: Potentials and challenges," *IEEE Communications Surveys & Tutorials*, vol. 19, no. 2, pp. 721–742, 2016.

[47] T. Takeda and K. Higuchi, "Enhanced user fairness using non-orthogonal access with SIC in cellular uplink," in *2011 IEEE vehicular technology conference (VTC Fall)*. IEEE, 2011, pp. 1–5.

[48] D. Caudill, P. B. Papazian, C. Gentile, J. Chuang, and N. Golmie, "Omnidirectional channel sounder with phased-arrayantennas for 5g mobile communications," *IEEE Transactions on Microwave Theory and Techniques*, vol. 67, no. 7, pp. 2936–2945, 2019.

[49] L. Zhang, Y. Wu, W. Li, K. Salehian, S. Lafleche, X. Wang, S. I. Park, H. M. Kim, J.-y. Lee, N. Hur *et al.*, "Layered-division multiplexing: An enabling technology for multicast/broadcast service delivery in 5G," *IEEE Communications Magazine*, vol. 56, no. 3, pp. 82–90, 2018.

Muhammad Kamran Naeem (M'17) received the PhD degree in Telecommunication Networks from Staffordshire University, Stafford, UK, 2017. He was a research fellow with Birmingham City University before joining Solent University, UK, as a postdoctoral researcher in 2018. He is now working as a postdoctoral research fellow at the University of Wolverhampton. His research interests include 6G networks, NOMA and internet of things (IoT) communications. Previously, he worked on topics including wireless sensor networks and channel estimation and equalization techniques.

Raouf Abozariba (S'16 - M'17) received the PhD degree from Staffordshire University, 2017, and he was a Senior Research Associate with the School of Computing and Communications at Lancaster University, UK. He joined Birmingham City University, UK, as a lecturer in Wireless Networking Technologies, 2019. His current research focuses on 5G RAN with an emphasis on NOMA, dynamic spectrum sharing and allocation.

Md Asaduzzaman (M'16 - SM'19) received the BSc and MSc degrees in applied statistics from the University of Dhaka, Dhaka, Bangladesh, in 1999 and 2001, respectively, the MSc degree in bioinformatics from the Chalmers University of Technology, Gothenburg, Sweden, in 2007, and the PhD degree in operational research from the University of Westminster, London, UK, in 2010. He is currently an Associate Professor in Operational Research with Staffordshire University, Stoke-on-Trent, UK, where he has been a Faculty Member, since 2014. His primary research interests include queuing, other stochastic and optimization models for performance measure, capacity and resource planning and management in communication networks. He is also interested in statistical computing, large-scale data mining, and analysis. Dr. Asaduzzaman received several awards, including the First Runner-Up Prize of The Doctoral Award from the Operational Research Society, UK, in 2011, and the Dean's Honour Award from the University of Dhaka, Bangladesh, in 2013.

Mohammad N. Patwary (SM'11) received the B.Eng. degree (Hons.) in electrical and electronic engineering from the Chittagong University of Engineering and Technology, Bangladesh, in 1998, and the PhD degree in telecommunication engineering from The University of New South Wales at Sydney, Sydney, NSW, Australia, in 2005. He was with General Electric Company, Bangladesh, from 1998 to 2000, and Southern Poro Communications, Sydney, from 2001 to 2002, as a Research and Development Engineer. He was a Lecturer with The University of New South Wales at Sydney, from 2005 to 2006, and then a Senior Lecturer with Staffordshire University, UK, from 2006 to 2010. He was then a Full Professor of wireless systems and digital productivity and the Chair of the Centre of Excellence on Digital Productivity with Connected Services, Staffordshire University, until 2016. He was a Full Professor of telecommunication networks and digital productivity and the Head of the Intelligent Systems and Networks (ISN) Research Group, School of Computing and Digital Technology, Birmingham City University, UK, from 2017 till 2020. He was research lead for a world's first '5G Connected Forest' project - accelerating destination branding for visitor economy in the UK. He is currently a full Professor of Telecommunications and Director of Centre for Future Networks & Autonomous Systems at the Faculty of Science & Engineering in The University of Wolverhampton, UK. His current research interests include - wireless communication networks & systems design and optimization, signal processing and energy-efficient systems, smart & autonomous systems, future generation of cellular network architecture and business modelling for Data-economy.