IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received March 20, 2021, accepted April 6, 2021, date of publication April 8, 2021, date of current version April 20, 2021.

Digital Object Identifier 10.1109/ACCESS.2021.3071992

User Clustering and Optimized Power Allocation for D2D Communications at mmWave Underlaying MIMO-NOMA Cellular Networks

SUHARE SOLAIMAN[1,2], LAILA NASSEF[1], AND ETIMAD FADEL[1], (Member, IEEE)
[1]Department of Computer Science, Faculty of Computing and Information Technology, King Abdulaziz University, Jeddah 21589, Saudi Arabia
[2]Department of Computer Science, College of Computers and Information Technology, Taif University, Taif 26571, Saudi Arabia

Corresponding author: Suhare Solaiman (ssuliman0007@stu.kau.edu.sa)

ABSTRACT Fifth-generation (5G) cellular networks are being developed to meet the ever-growing data traffic across mobile devices and their applications. The core of 5G cellular networks is leveraging wider and higher frequencies available at millimeter wave frequency (mmWave) bands, thus providing very high data rates for mobile devices. Multi-input multi-output (MIMO) is an essential technology for overcoming the high propagation loss at mmWave communications. In non-orthogonal multiple access (NOMA), multiple cellular user equipments (CUEs) communicate over the same time-frequency resources using a multiplexed power domain. In device-to-device (D2D) communications, two D2D user equipments (DUEs) communicate without passing through the base station. In the underlaying scenario, DUEs reuse the frequency resources allocated to CUEs for spectrum utilization but DUEs cause interferences for cellular and D2D communications. Integrating D2D communications with other 5G technologies has great potential for spectral efficiency improvement. Unfortunately, interference management and resource allocation are becoming increasingly challenging due to aggressive frequency reuse. In this paper, D2D communications at mmWave underlaying MIMO-NOMA cellular network system model is developed. Consequently, a novel resource allocation for D2D communications underlaying MIMO-NOMA cellular network is proposed. A resource allocation optimization problem is formulated for spectral efficiency maximization. To solve this NP-hard problem, the problem is decomposed into three subproblems: interference-aware graph-based user clustering, MIMO-NOMA beamforming design, and optimized power allocation based on particle swarm optimization. Simulation results demonstrate that the proposed algorithm for D2D communications at mmWave underlaying MIMO-NOMA cellular network delivers a greater spectral efficiency compared to the conventional D2D communications that operate underlay MIMO-orthogonal multiple access cellular networks.

INDEX TERMS Device-to-device communication, interference management, millimeter wave communication, MIMO, NOMA, power allocation, resource allocation.

I. INTRODUCTION

Currently, the massive growth in the number of mobile devices and their high-speed applications has accelerated the ever-growing flow of mobile data traffic. Monthly global mobile data traffic will reach 77 exabytes per month in 2022, and this trend will continue, as estimated by Cisco [1]. As this growth occurs, the demands for high data rate, low latency, and highly reliable wireless communications are dramatically increasing [2]. Fifth-generation (5G) cellular networks have promised to satisfy these demands for future applications and services. The key technologies required for enabling 5G cellular networks essentially include millimeter wave frequency (mmWave) communications, multi-input multi-output (MIMO), non-orthogonal multiple access (NOMA), and device-to-device (D2D) communications. The microwave frequency bands have become highly congested and cannot accommodate the exponential increase in communication capacity due to the limited frequency resources [3]. For this reason, mmWave is becoming the core of 5G cellular networks. mmWave leverages a wider bandwidth at high frequencies (ranging from 24 to 300 GHz) [4], thereby offering enormous amount of bandwidth that can be utilized not only to accommodate increased

The associate editor coordinating the review of this manuscript and approving it for publication was Hosam El-Ocla.

communication capacity but also to satisfy the increasing demand for high data rates communications. Different from microwave propagation characteristics, the mmWave signals are prone to high path-loss propagation [3], [5]. Fortunately, the short wavelengths of mmWave signals allow large-scale antennas to be positioned in a limited physical area, and this can be obtained by MIMO technology [6], [7]. These directional antennas are capable of transmitting or receiving signals through beamforming techniques in specific directions [6], thereby offering high beamforming and spatial processing gains that can overcome the high path-loss propagation [7]. Another way to overcome the high path-loss is the deployment of large-scale small cells in an urban environment over a coverage range of approximately 150-200 meters, as demonstrated by recent channel measurements [8].

In MIMO, the maximum number of multiplexed data streams that can be transmitted simultaneously over a wireless channel is determined by the number of Radio Frequency (RF) chains. The spectral efficiency is thus proportional to the number of multiplexed data streams [9]. Digital beamforming is a well-developed technique for conventional MIMO, where the number of RF chains is equal to the number of antennas, and these systems are equipped with few antennas (approximately 10) [7], [10]. However, the use of large-scale antennas in mmWave networks tends to result in an equally large number of RF chains involving higher hardware and energy consumption [3]; therefore, it is not practical to use digital beamforming for mmWave networks. Furthermore, analog beamforming has only one RF chain to serve the transmission of a single data stream and is therefore not feasible for multi-user or multi-stream scenarios [8]. Analog beamforming is supported by phase shifters that control the signal phase at each antenna [8]. It is currently a de-facto solution for indoor mmWave networks [11]. In contrast, hybrid beamforming (HB) is recently proposed for outdoor mmWave networks by enabling mmWave-MIMO to achieve the performance of MIMO digital beamforming with fewer RF chains [10].

In downlink power-domain NOMA, different cellular user equipments (CUEs) are assigned different power levels according to their channel gains, whereas multiple CUEs reuse the same time-frequency resources [12]. Thus, NOMA is beneficial for supporting a huge number of CUEs in spectrally efficient communications [13]. At the base station (BS), NOMA invokes the superposition coding for multiple streams and then transmits the superimposed signal over the same time-frequency resources via power multiplexing [14], [15]. Different CUEs as receivers adopt the successive interference cancellation (SIC) technique to eliminate the intra-beam interference and recover their desired signals [14]. In contrast, each CUE is delegated to orthogonal resources in the time-domain, frequency-domain, code-domain, or their combinations in conventional orthogonal multiple access (OMA) [12]. In MIMO-NOMA beamforming, multiple transmission antennas at the BS are employed to generate various beams in the spatial domain where each beam adopts the fundamental NOMA technology [14]. This involves designing multi-user beamforming with a single beamforming vector to support multiple CUEs in a NOMA cluster. The number of RF chains in MIMO-NOMA beamforming is reduced, where each beam is served by a single RF chain. In MIMO-OMA beamforming, only one CUE is served by each beam, which is orthogonal to other beams in terms of frequency. This leads to increased hardware and energy consumption since the maximum number of CUEs that the BS can serve simultaneously is equal to the number of RF chains, and the number of beams cannot exceed the number of RF chains [16]. Due to the high demand for the bandwidth needed to support large-scale users with high data rate communications and lower energy consumption, employing MIMO-NOMA at mmWave becomes a natural choice for 5G cellular networks.

D2D communication is a direct link between two nearby D2D user equipments (DUEs) without transmitting data via a BS [17]. In the D2D communications underlaying cellular network, DUEs are allowed to reuse the licensed frequencies allocated to CUEs for spectrum utilization, but in the meantime they cause interference to both cellular communications and other D2D communications. Therefore, careful interference management and resource allocation are needed in the underlaying scenario to improve the performance of the network. D2D communications promise significant improvements to the cellular network [18], [19]. Due to the direct communication between DUEs at low levels of transmitted power, D2D communications can provide ultra-low latency, increased data rate, offloaded BS traffic, and reduced energy consumption. In addition, D2D communications can reuse the same frequency resources allocated to CUEs, thereby facilitating dense spectral utilization and improving the spectral efficiency of the network. Furthermore, D2D communications can extend the coverage of the current cellular network without additional infrastructure expenses. The technology of D2D communications has promised to launch several new proximity-based applications and services into cellular networks, such as public safety, social and commercial services, coverage extension, BS data and computation offloading, and vehicle-to-vehicle communications [20].

A. MOTIVATION

Although large bandwidth is available at mmWave bands, the number of RF chains is limited since they cause high hardware and energy consumption at mmWave bands [21]. In such a situation, the number of CUEs that can be served under one resource block is no greater than the RF chains [21]. To overcome this limitation and increase the number of CUEs, NOMA is necessary to be implemented into mmWave communications. In addition, the integration of D2D communications with MIMO-NOMA at mmWave is required to provide services capable of handling the data streaming connection of the expected large number of connected devices in dense networks. This integration is capable of offloading significant pressure on BS and utilizes the available spectrum by providing proximity-based applications and services.

One potential application scenario of D2D communications at mmWave underlaying MIMO-NOMA cellular network is serving density live content streaming, such as in a stadium. The massive traffic flows place tremendous pressure on the BS and spectrum resources. Thus, D2D communications can provide media servers that deliver media services in D2D mode to large number of DUEs. In addition, DUEs can use D2D communications to get the media content from nearby DUEs that have acquired media content services. In this way, the downlink transmission pressure of cellular network BS can be offloaded. This has motivated the integration of D2D communications at mmWave underlaying MIMO-NOMA cellular network to improve the spectral efficiency and energy efficiency of the network by utilizing the least number of resource blocks under certain constraints to serve all CUEs and DUEs in the coverage area.

B. RELATED WORK

Although comprehensive studies have been conducted on resource allocation in cellular networks, studies on the integration of D2D communications at mmWave with MIMO-NOMA are very limited.

A game-based interference management algorithm for D2D communications underlaying mmWave small cell network was developed by the authors in [22]. The algorithm was developed to optimize D2D communications power allocation and to minimize interference caused by D2D communications while taking full advantage of mmWave bandwidth. Their simulation results showed that the proposed algorithm converged rapidly, retained a high range of signal-to-interference-plus-noise ratios (SINRs), and obtained excellent throughput performance.

In a downlink cellular network with underlaying D2D communications, the authors in [23] proposed beamforming based multi-user MIMO-NOMA. Two algorithms for multi-user MIMO beamforming were developed. The first algorithm was intended to eliminate the inter-beam interference, while the second was intended to eliminate the interference from the BS to D2D communications. To maximize the sum throughput, an optimization problem was formulated. A suboptimal sequential algorithm was developed to solve this non-deterministic polynomial-time (NP)-hard problem by designing a zero-forcing (ZF) beamforming matrix for multi-user MIMO and then employing user grouping and an optimal power allocation algorithm. The simulation results showed that the integration of multi-user MIMO beamforming, NOMA, and D2D communications improved significantly the throughput of the network.

In [2], the authors studied the outage probability and the ergodic capacity in downlink MIMO-NOMA at mmWave cellular network with D2D communications. Analytical results indicated that NOMA outperforms time division multiple access (TDMA). The performance factors, including the transmission power and the number of antennas were analyzed. Higher transmission power and more antennas in the BS have been found to decrease the outage probability and enhance the ergodic capacity of NOMA.

The authors in [24] considered cellular and D2D communications underlaying NOMA. They aimed to provide interference management against inter-cluster and intra-cluster interferences. In addition, they provided algorithms for optimized user clustering and power allocation with the objectives of maximizing the sum-rate of the network. In terms of the average sum-rate, simulation results revealed that the proposed algorithm achieved up to 70% and 92% of gains compared with the fundamental NOMA and conventional orthogonal frequency division multiple access (OFDMA), respectively. Furthermore, the results showed that in terms of the number of admitted users, the proposed algorithm significantly increased network connectivity. In addition, the integration of D2D communications with cloud-RAN networks was considered in [25] to solve the delay issue caused by increasing mobile traffic.

The authors in [10] and [26] investigated the application of NOMA at mmWave with downlink a HB architecture. In particular, a user grouping algorithm was first proposed according to the user channel correlations with the aim of mitigating the interference between different clusters. Then, a joint HB and power allocation problem was formulated to maximize the spectral efficiency [10] in sum-rate in [26]. They adopted ZF to eliminate the inter-beam interference. In addition, they used SIC to eliminate the intra-beam interference. Simulation results showed that the mmWave NOMA system surpasses mmWave OMA in terms of sum-rate, spectral efficiency, and energy efficiency. Furthermore, the proposed multi-beam NOMA scheme at mmWave in [27] provided a greater flexibility in serving multiple CUEs compared to the conventional single-beam mmWave-NOMA scheme. In addition, mmWave-NOMA systems were considered in [28] to achieve high data rates in 5G ultra dense networks.

In [29], the authors proposed a beamwidth control technique to increase the number of admitted NOMA clusters supported by widening the beamwidth that can further exploit the energy efficiency gain of NOMA. A joint user pairing and power allocation problem was considered in [30] to optimize the achievable sum-rate in a downlink NOMA network and a joint beamforming and power allocation for mmWave-NOMA network was proposed in [31].

C. PROBLEM STATEMENT

An illustration of D2D communications underlaying 5G macrocell cellular network is shown in Fig. 1, the macrocell is divided into small cells (i.e., microcells). The BSs are placed at the centers of small cells and connected to the macrocell BS via gateways. Each small cell BS serves cellular and D2D communications at mmWave underlaying MIMO-NOMA cellular network. In this model, sets of CUEs are grouped into clusters, where different clusters within the same BS are assigned orthogonal frequency resources, and each cluster is served by a beam with a single RF chain. By using power-domain NOMA, CUEs in the same cluster operate within the

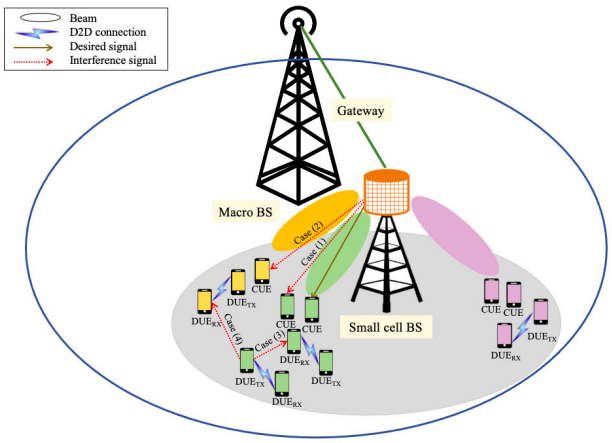

FIGURE 1. Illustration of D2D communications underlaying 5G macrocell cellular network.

same time-frequency mmWave resource block with different power coefficients. For spectrum utilization, multiple D2D pairs (i.e., DUE$_{Tx}$ as transmitter and DUE$_{Rx}$ as receiver) are matched with CUE clusters, where D2D pairs are allowed to reuse the same frequency resources allocated to CUEs in the same cluster. For energy efficiency purposes, MIMO-NOMA HB is designed for the BS to steer the beams in specific directions with a limited number of RF chains.

The main challenge of this model is the existence of great interference due to aggressive frequency reuse. As shown in Fig. 1, downlink D2D communications in mmWave underlaying MIMO-NOMA cellular network cause four cases of interference:

- Case (1) is intra-beam interference, where the signal transmitted by a beam from a BS to a CUE causes interference to other CUEs served by this beam.
- Case (2) is inter-beam interference, where the signal transmitted by a beam from a BS to a CUE causes interference to other CUEs served by adjacent beams.
- Case (3) is intra-cluster interference, where a DUE$_{Tx}$ causes interference to CUEs and DUE$_{Rx}$s within the same cluster.
- Case (4) is inter-cluster interference, where a DUE$_{Tx}$ causes interference to CUEs and DUE$_{Rx}$s in different clusters.

However, it is highly desired to mathematically develop a system model for D2D communications at mmWave underlaying MIMO-NOMA cellular network. Consequently, proposing a novel resource allocation design for intelligent management and mitigation of various interference cases under all above-mentioned network assumptions and capabilities. The joint integration of D2D communications at mmWave underlaying MIMO-NOMA cellular network is considered for the first time in this paper, and these four interference cases are also examined and managed jointly for the first time. Optimizing the resource allocation problem to maximize the spectral efficiency while guaranteeing the quality-of-service (QoS) of both CUEs and D2D pairs and providing interference protection for CUEs and DUE$_{Rx}$s results in non-convex mixed-integer non-linear programming (MINLP) problem. A non-convex MINLP problem is considered to be NP-hard and is not appropriate for realistic real-time implementation. To solve this problem, the resource allocation problem is decomposed into three subproblems: interference-aware graph-based user clustering, MIMO-NOMA HB design, and optimized power allocation based on particle swarm optimization (PSO).

D. CONTRIBUTIONS

The contributions of this paper are summarized as follows:

- A system model for single-cell downlink D2D communications at mmWave underlaying MIMO-NOMA cellular network is developed. In this model, CUEs are grouped into clusters and MIMO-NOMA beamforming signals are transmitted to these clusters. Moreover, D2D pair clusters are matched to CUE clusters, where D2D pairs are allowed to reuse the same mmWave resource blocks occupied by CUEs in the same cluster.
- The resource allocation optimization problem for D2D communications underlaying downlink MIMO-NOMA cellular network is formulated. The objective of this problem is to maximize the spectral efficiency of the network while guaranteeing the QoS of both CUEs and D2D pairs and providing interference protection for CUEs and DUE$_{Rx}$s. The formulated problem belongs to the non-convex MINLP class. Therefore, it is incredibly difficult to obtain an optimal global solution to the proposed problem. To solve this problem, it is decomposed into three subproblems: user clustering, beamforming design, and power allocation.
- A novel interference-aware graph-based user clustering algorithm is proposed to solve the user clustering problem, where the algorithm defines the best cluster of CUEs for MIMO-NOMA HB and the best user cluster (i.e., CUEs and D2D pairs) for spectrum sharing.
- A MIMO-NOMA HB is designed with limited RF chains to increase the energy efficiency, where a single RF chain is required for each beam and a single beam serves a cluster of CUEs.
- A novel optimized power allocation algorithm based on PSO is proposed.
- Simulation results demonstrate that the proposed resource allocation algorithm for D2D communications at mmWave underlaying MIMO-NOMA cellular network delivers a greater spectrum efficiency and energy efficiency compared to the conventional D2D communications that operate underlay MIMO-OMA cellular network.

E. PAPER ORGANIZATION

This paper is organized as follows. Section II presents the system model of D2D communications at mmWave underlaying MIMO-NOMA cellular network and the problem

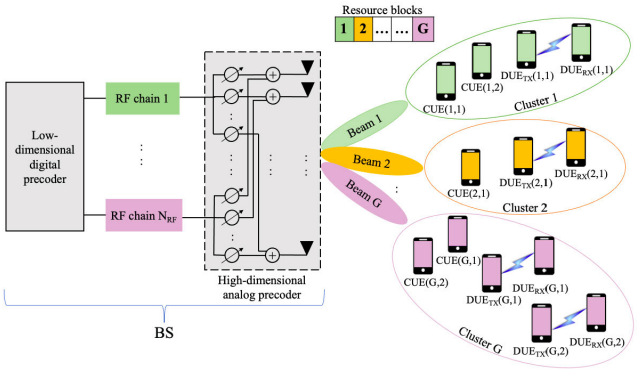

FIGURE 2. The system model of D2D communications at mmWave underlaying MIMO-NOMA cellular network.

formulation. In Section III, the interference-aware graph-based user clustering is proposed. In Section IV, the designed MIMO-NOM HB is presented. In Section V, the optimized power allocation based on PSO is proposed. Section VI shows the simulation results and performance analysis. Finally, the conclusion and future work are discussed in Section VII.

II. SYSTEM MODEL AND PROBLEM FORMULATION

The system model of D2D communications at mmWave underlaying a single small-cell downlink MIMO-NOMA cellular network is developed, as shown in Fig. 2. There are Q CUEs distributed randomly, each with one antenna. Additionally, there are Z D2D pairs randomly distributed. A DUE_{Rx} has a single receiver antenna and receives a directional antenna array composed of N_D antennas from a DUE_{Tx} supported by an independent RF chain. The BS is equipped with N_{BS} antennas and N_{RF} RF chains that can generate G high directional beams to simultaneously serve G clusters [10], [32], where $N_{RF} < Q + Z < N_{BS}$. In particular, the number of G beams (i.e., the number of G clusters) is equal to the number of N_{RF} chains, where $G = N_{RF}$ and the G is not larger than N_{RF}. Thus, Q CUEs and Z D2D pairs should be partitioned into G clusters, and each cluster should be supported by an independent RF chain.

The channels for all communication links are estimated by the BS as follows:

A. NOMA SIGNAL

In beam g, NOMA allows a set of K CUEs to be scheduled on the same time-frequency mmWave resource block. The set of K CUEs served by beam g is denoted as $S_g = \{CUE(g,1), CUE(g,2), \ldots, CUE(g,k)\}$ for $g = 1, 2, \ldots, G$. Here, $CUE(g,k)$ denotes that the CUE is served by the g-th beam with sequence k in that beam, where $|S_g| \geq 1$ and $S_i \cap S_j = \phi$ for $i \neq j$. According to NOMA, the BS transmits a superimposed signal X_g for all K CUEs in the g-th cluster via g-th beam [2], [23], and X_g can be expressed as

$$X_g = \sum_{k=1}^{K} \sqrt{\beta_{CUE(g,k)} P_g} S_{CUE(g,k)}, \quad (1)$$

where $S_{CUE(g,k)}$ is the transmitted signal. $\beta_{CUE(g,k)}$ is the power coefficient allocated to $CUE(g,k)$, where $\sum_{K}^{k=1} \sqrt{\beta_{CUE(g,k)} P_g} = 1$. P_g is the total transmitted power for the g-th beam. Without loss of generality, the transmission power is assumed to be equally divided between G beams, i.e., $P_g = \frac{P_{BS}}{N_{BS}}$, where P_{BS} is the BS total transmission power provided by the BS.

B. mmWave CHANNEL MODEL

The widely used directional mmWave channel model with L scatters and a uniform linear array (ULA) with a half-wavelength antenna spacing proposed in [33] is adopted. Under this model, the channel vector $N_{BS} \times 1$ of $CUE(g,k)$ is denoted as $H_{CUE(g,k)}$ and can be expressed as

$$H_{CUE(g,k)} = \sqrt{\frac{N_{BS}}{PL_{CUE(g,k)}}} \sum_{l=1}^{L_{g,k}} \alpha_{g,k}^{(l)} a(\varphi_{g,k}^{(l)}) a(\theta_{g,k}^{(l)}), \quad (2)$$

where $L_{(g,k)}$ represents the number of paths between the BS and $CUE(g,k)$; $\alpha_{g,k}^{(l)}$ is the complex gain of path l; $PL_{CUE(g,k)}$ represents the average path-loss between the BS and $CUE(g,k)$; $\varphi_{(g,k)}$ and $\theta_{(g,k)} \in [0, 2\pi]$ are angle-of-arrival (AoA) and angle-of-departure (AoD) of path l, respectively; and $a(\varphi_{g,k}^{(l)})$ and $a(\theta_{g,k}^{(l)})$ represent the BS antenna steering vector and $CUE(g,k)$ antenna responding vector, respectively. The steering vector $a(\varphi_{g,k}^{(l)})$ of $N_{BS} \times 1$ can be defined as

$$a(\varphi_{g,k}^{(l)}) = \frac{1}{\sqrt{N_{BS}}} [1, e^{j(\frac{2\pi}{\lambda})\Omega \sin(\varphi)}, \ldots, e^{j(N_{BS}-1)(\frac{2\pi}{\lambda})\Omega \sin(\varphi)}]^T \quad (3)$$

where λ is the signal wavelength, Ω is the distance between antennas, and j is an imaginary element. The response vector at a given CUE, $a(\theta_{g,k}^{(l)})$ can be defined using a similar equation to (3).

The same channel model in (2) is adopted for D2D communication, and the channel between a D2D pair in the g-th beam is denoted as $H_{DUE(g,m)}$ and expressed as

$$H_{DUE(g,m)} = \sqrt{\frac{N_D}{PL_{DUE(g,m)}}} \sum_{l=1}^{L_{g,m}} \alpha_{g,m}^{(l)} a(\varphi_{g,m}^{(l)}) a(\theta_{g,m}^{(l)}), \quad (4)$$

where N_D is the number of DUE_{Tx} antennas; $L_{g,m}$ represents the number of paths between the m-th DUE_{Tx} and DUE_{Rx}; $\alpha_{g,m}^{(l)}$ is the complex gain of path l; $PL_{DUE(g,m)}$ represents the average path-loss between the m-th DUE_{Tx} and DUE_{Rx}; $\varphi_{(g,m)}$ and $\theta_{(g,m)} \in [0, 2\pi]$ are the AoA and AoD of path l, respectively; $a(\varphi_{g,m}^{(l)})$ and $a(\theta_{g,m}^{(l)})$ are the antenna steering vector of DUE_{Tx} and antenna responding vector of DUE_{Rx}, respectively. They can be defined using a similar equation to (3).

Based on the close-in path-loss model [34], the PL for $CUE(g,k)$ over a given frequency and distance can be expressed as

$$PL_{CUE(g,k)}(f, dist) = 20 \log(4\pi \frac{f}{\varsigma}) + 10\delta log(dist), \quad (5)$$

where *dist* is the transmitter-receiver distance, f is the carrier frequency, ς is the speed of light, and δ is the path-loss exponent. Both line of sight (LOS) and non-line of sight (NLOS) are considered where $PL \in \{LOS, NLOS\}$. In addition, the same model in (5) is used to determine the PL for $DUE(g, m)$.

C. RECEIVED SIGNAL

The CUEs receive the superimposed signal from the BS and along with other interference signals. These interference signals include case (1) (intra-beam), case (2) (inter-beam), case (3) (intra-cluster), and case (4) (inter-cluster) signals, as shown in Fig. 1 and denoted as $I_{CUE(g,\acute{k})}$, $I_{CUE(\acute{g},\acute{k})}$, $I_{DUE(g,m)}$, and $I_{DUE(\acute{g},m)}$, respectively.

The signal received by $CUE(g, k)$ can be formulated as

$$\begin{aligned} Y_{CUE(g,k)} = {} & H_{CUE(g,k)} A d_g \sqrt{\beta_{CUE(g,k)} P_g} S_{CUE(g,k)} \\ & + I_{CUE(g,\acute{k})} + I_{CUE(\acute{g},\acute{k})} + I_{DUE(g,m)} \\ & + I_{DUE(\acute{g},m)} + \eta_{CUE(g,k)}, \end{aligned} \tag{6}$$

where

$$I_{CUE(g,\acute{k})} = H_{CUE(g,k)} A d_g \sum_{\acute{k}=1,\acute{k}\neq k}^{K} \sqrt{\beta_{CUE(g,\acute{k})} p_g} \times S_{CUE(g,\acute{k})},$$

$$I_{CUE(\acute{g},\acute{k})} = H_{CUE(g,k)} A \sum_{\acute{g}=1,\acute{g}\neq g}^{G} d_{\acute{g}} \sum_{\acute{k}=1}^{K} \sqrt{\beta_{CUE(\acute{g},\acute{k})} p_{\acute{g}}} \times S_{CUE(\acute{g},\acute{k})},$$

$$I_{DUE(g,m)} = \sum_{m=1}^{M} \sqrt{P_{DUE(g,m)}} H_{DUE(g,m),CUE(g,k)} \times S_{DUE(g,m)},$$

$$I_{DUE(\acute{g},m)} = \sum_{\acute{g}=1,\acute{g}\neq g}^{G} \sum_{m=1}^{M} \sqrt{P_{DUE(\acute{g},m)}} H_{DUE(\acute{g},m),CUE(g,k)} \times S_{DUE(\acute{g},m)},$$

$H_{CUE(g,k)}$ is the channel gain of $CUE(g, k)$, A of size $N_{BS} \times N_{RF}$ is the analog precoding matrix, d_g of size $N_{RF} \times 1$ is the digital precoding vector for the g-th beam, $I_{CUE(g,\acute{k})}$ is the intra-beam interference caused by the BS, $I_{CUE(\acute{g},\acute{k})}$ is the inter-beam interference caused by the BS, $I_{DUE(g,m)}$ is the intra-cluster interference caused by $DUE_{Tx}(g, m)$, $I_{DUE(\acute{g},m)}$ is the inter-cluster interference caused by $DUE_{Tx}(\acute{g}, m)$, and $\eta_{CUE(g,k)}$ is the additive white Gaussian noise.

The signal received by $DUE_{Rx(g,m)}$ can be formulated as

$$\begin{aligned} Y_{DUE(g,m)} = {} & H_{DUE(g,m)DUE(g,m)} A d_{DUE(g,m)} \sqrt{P_{DUE(g,m)}} \\ & \times S_{DUE(g,m)} + I_{DUE(g,\acute{m})} + I_{DUE(\acute{g},\acute{m})} \\ & + \eta_{DUE(g,m)}, \end{aligned} \tag{7}$$

where

$$I_{DUE(g,\acute{m})} = \sum_{\acute{m}=1,\acute{m}\neq m}^{M} H_{DUE(g,\acute{m})DUE(g,m)} \sqrt{P_{DUE(g,\acute{m})}} \times S_{DUE(g,\acute{m})},$$

$$I_{DUE(\acute{g},\acute{m})} = \sum_{\acute{g}=1,\acute{(g)}\neq g}^{G} \sum_{\acute{m}=1}^{M} H_{DUE(\acute{g},\acute{m})DUE(g,m)} \times \sqrt{P_{DUE(\acute{g},\acute{m})}} S_{DUE(\acute{g},\acute{m})},$$

$H_{DUE(g,m)}$ is the channel gain from $DUE_{Tx}(g, m)$ to $DUE_{Rx}(g, m)$, S is the transmitted signal, A of size $N_D \times 1$ is D2D analog precoder; $d_{DUE(g,m)}$ is D2D digital precoder, $I_{DUE(g,\acute{m})}$ is the intra-cluster interference caused to $DUE_{Rx}(g, m)$ by different $DUE_{Tx}(g, \acute{m})$, $I_{DUE(\acute{g},\acute{m})}$ is the inter-cluster interference caused to $DUE_{Rx}(g, m)$ by different $DUE_{Tx}(\acute{g}, \acute{m})$, P_{DUE} is the transmitted power for D2D pair, and $\eta_{DUE(g,m)}$ is the additive white Gaussian noise.

D. SUM-RATE FORMULATION

The SINRs of $CUE(g, k)$ and $DUE_{Rx}(g, m)$ can be formulated as

$$\gamma_{CUE(g,k)} = \frac{\sqrt{\beta_{CUE(g,k)P_g}} ||\bar{H}_{CUE(g,k)} d_g||^2}{I_{CUE(g,\acute{k})} + I_{CUE(\acute{g},\acute{k})} + I_{DUE(g,m)} + I_{DUE(\acute{g},m)} + \eta_{CUE(g,k)}}, \tag{8}$$

$$\gamma_{DUE(g,m)} = \frac{\sqrt{P_{DUE(g,m)}} ||\bar{H}_{DUE(g,m)DUE(g,m)} d_{DUE(g,m)}||^2}{I_{DUE(g,\acute{m})} + I_{DUE(\acute{g},\acute{m})} + \eta_{DUE(g,m)}}, \tag{9}$$

where $\bar{H}$ is the equivalent channel gain vector. Given SINRs, the data rates for $CUE(g, k)$ and $DUE_{Rx}(g, m)$ can be calculated by using the Shannon capacity formula as follow:

$$R_{CUE(g,k)} = \log_2(1 + \gamma_{CUE(g,k)}), \tag{10}$$

$$R_{DUE(g,m)} = \log_2(1 + \gamma_{DUE(g,m)}). \tag{11}$$

The total sum-rate can be expressed as

$$R = \sum_{g=1}^{G} (\sum_{k=1}^{K} R_{CUE(g,k)} + \sum_{m=1}^{M} R_{DUE(g,m)}). \tag{12}$$

E. PROBLEM FORMULATION

In this paper, the spectrum efficiency is determined as the total sum-rate of both CUEs and DUE_{Tx}s as defined in (12). The resource allocation optimization problem of D2D communications at mmWave underlaying MIMO-NOMA cellular network with the objective of maximizing the spectral efficiency is formulated as

$$\max(R), \tag{13}$$

Subject to

$$\begin{aligned} C1: \quad & \gamma_{CUE(g,k)} \geq \gamma_{CUE(g,k),min}, \\ C2: \quad & \gamma_{DUE(g,m)} \geq \gamma_{DUE(g,m),min}, \\ C3: \quad & \textstyle\sum_{g=1}^{G} \sum_{k=1}^{K} \beta_{CUE(g,k)} P_g \leq P_{BS_{max}}, \\ C4: \quad & \textstyle\sum_{g=1}^{G} \sum_{m=1}^{M} P_{DUE(g,m)} H_{(DUE(g,m),CUE(g,k))} \leq th^g, \\ C5: \quad & 0 < P_{CUE(g,k)} < P_{CUE_{max}}, \\ C6: \quad & 0 < P_{DUE(g,m)} < P_{DUE_{max}}, \end{aligned}$$

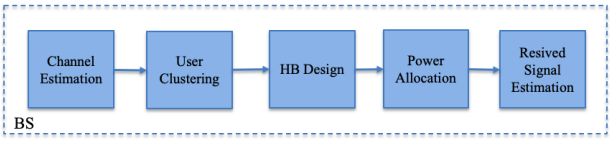

FIGURE 3. The framework of the proposed resource allocation algorithm.

where $\gamma_{CUE,min}$ and $\gamma_{DUE,min}$ are the minimum SINRs required to satisfy the QoS of CUEs and D2D pairs, respectively. $C1$ and $C2$ guarantee minimum SINRs of CUEs and D2D pairs, respectively. For cost-effective energy consumption, $C3$ indicates the power limit assigned to the CUEs, with $P_{BS_{max}}$ being the maximum transmitted power of the BS. For intra-cluster mitigation, $C4$ guarantees that interference caused to $CUE(g,k)$ by different $DUE_{Tx(g,m)}$ in the g-th cluster does not exceed the interference threshold th^g for $\forall g \in G$. Under this constraint, a set of D2D pairs is allowed to be scheduled with the CUE resources only if the interference protection can be guaranteed for the CUEs. $C5$ and $C6$ indicate the power allocation limits for each CUE and D2D pair, respectively.

The optimization problem in (13) is non-convex due to the non-convexity of the objective function, and it belongs to the MINLP class since a combination of integer and continuous numbers are used in the optimization variables. A non-convex MINLP problem is rarely solved by theoretical analysis. However, the proposed solution to solve the optimized problem in (13) is decomposed into three subproblems. First, interference-aware user clustering based on graph theory is proposed to provide a suboptimal user clustering solution. Second, MIMO-NOMA HB at mmWave is designed to achieve efficient spatial multiplexing despite the limited number of RF chains. Third, an optimized power allocation algorithm based on PSO is proposed to maximize the network spectral efficiency.

Fig. 3 shows the framework of the proposed resource allocation algorithm. Based on the use of time division duplex (TDD) as spectrum usage technique and the full knowledge of channel state information (CSI), CUEs and D2D pairs are sending the CSIs to the BS in uplink transmission, then the BS estimates the channel of all communications as described in Section II. Then, the BS performs user clustering for CUEs and D2D pairs as presented in Section III. Then, the BS designs the HB with digital precoder and analog precoder as described in Section IV. Then, the BS optimizes the power allocation for CUEs and D2D pairs in Section V. According to the value of channel estimation, digital precoder, analog precoder, and optimized power allocation the received signal is estimated by the BS as described in Subsection II-C. Then, the CUEs receive the downlink data transmission and decode its desired signal by using SIC since superposition transmission is utilized within a NOMA cluster. In addition, after the received signal is estimated, the DUE_{Tx}s start sending the data stream to its pair DUE_{Rx}s.

III. INTERFERENCE-AWARE GRAPH-BASED USER CLUSTERING

As seen in the previous section, the formulated problem in (13) is a NP-hard problem whose solutions are combinatorial by nature. Particularly, for spectral efficiency maximization, the optimal solution for user clustering requires an exhaustive search to form a NOMA cluster [35]. This means that all possible combinations of user clustering must be considered for every single CUE [36]. In this context, the number of possible combinations of optimal NOMA user clustering for Q CUEs can be expressed as follows, $O(\sum_{i=2}^{Q}(\binom{Q}{i}))$ [36]. This time complexity will be further increased as D2D pair clustering and clusters matching are processed as the same way.

In this paper, user clustering is the foundation stone for resource allocation since it can improve network performance. Specifically, user clustering identifies the best user clusters (i.e., CUEs and D2D pairs) for spectrum utilization with minimal interference. In addition, it identifies the best cluster of CUEs for MIMO-NOMA beamforming, thus enhancing the beamforming gain by aligning the directional beam with a specific cluster. Furthermore, it facilitates the design of HB with cost-effective hardware, where each cluster is served by a single RF chain. Finally, it reduces the SIC process, where SIC invokes low-density CUEs in each cluster instead invoking all the CUEs served by the BS.

The interference awareness is a condition in which the BS gains local awareness of all communication links' channels and allowing this information to be used when allocating frequency resources to cellular and D2D communications. The interference awareness is achieved by applying the following procedures. First, the BS and DUE_{Tx}s (as transmitters) broadcast sounding signals periodically. Second, the CUEs and DUE_{Rx}s (as receivers) receive the sounding signals from the transmitters if they fall within the transmission coverage area; then, they measure the channel properties of their communications and interferences. Third, each receiver reports the CSI that they have collected to the BS. The receiver reports the channel gain as 0 if the receiver is outside the transmission coverage area. Finally, the BS collects all CSIs reported by the transmitters and therefore becomes the operator for cellular and D2D communications with regard to interference-aware scheduling.

Graph theory provides efficient tools to model and analyze many forms of connections, relationships, and processes across various networks, and has therefore been widely used to address the problems of resource allocation and interference management [37]. In this paper, graph theory is exploited for user clustering to provide a suboptimal interference-aware user clustering. Generally, graph-based clustering relies on partitioning the whole graph into disjoint subsets of related vertices in the graph [38]. The proposed user clustering algorithm consists of three main phases: CUE clustering, D2D pair clustering, and clusters matching. Here,

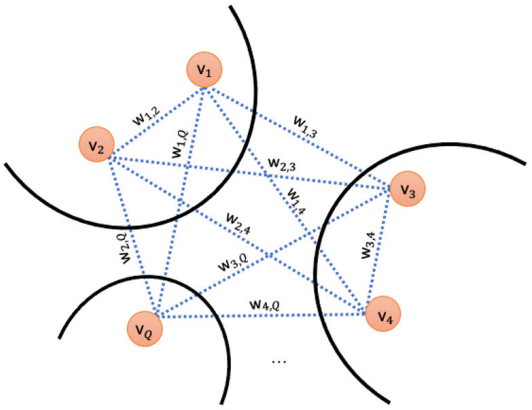

FIGURE 4. The channel correlation graph of CUEs.

the power coefficients of each CUEs and D2D pairs are assumed to be fixed, and the CSI is known perfectly.

A. CUE CLUSTERING

1) GRAPH CONSTRUCTION

The CUE channel graph is constructed by the BS and stated as problem P1 as follows:

P1: A graph $T_1 = (V, E)$ is given, where $V = \{\mathsf{v}_1, \ldots, \mathsf{v}_Q\}$, each vertex from set V represents a CUE, and each edge from E represents a weight. The edge weight represents the channel correlation between two CUEs. An edge weight between v_i,v_j is denoted by $\mathsf{w}_{i,j}$, as shown in Fig. 4. When there is no channel correlation between v_i and v_j, the edge weight $\mathsf{w}_{i,j} = 0$.

The goal of P1 is to partition graph T_1 into G disjoint clusters $\{C_1, \ldots, C_G\}$, where CUEs among clusters have low channel correlations to eliminate inter-beam interference and CUEs within a cluster have high channel correlations to improve the robustness of beamforming, as highly correlated channels have high beamforming and spatial multiplexing gains. Here, the edge weight between v_i and v_j represents the normalized channel correlation between v_i and v_j, which is expressed as

$$\mathsf{w}_{i,j} = \frac{H_i H_j}{\|H_i\| \|H_j\|}, \tag{14}$$

where H is defined as the same equation used in (2).

2) ALGORITHM DESIGNING

The most appropriate graph theory objective function that captures P1 is based on a multi-way normalized cut (Ncut) [39], which is formulated as

$$Ncut\{C_1, \ldots, C_G\} = \sum_{g=1}^{G} \frac{cut(C_g, \bar{C}_g)}{vol(C_g)}, \tag{15}$$

where the numerator $cut(C_g, \bar{C}_g)$ measures the similarity between cluster C_g and other clusters $\bar{C}_g$, and $cut(C_g, \bar{C}_g) = \sum_{\mathsf{v}_i \in C, \mathsf{v}_j \in \bar{C}} \mathcal{W}_{i,j}$. The denominator $vol(C_g)$ measures the similarity within clusters, where $vol(C_g) = \sum_{i \in C_g} \mathfrak{d}_{ii}$ and $\mathfrak{d}_{ii} = \sum_{j=1}^{Q} \mathsf{w}_{i,j}$.

TABLE 1. Graph-based CUEs clustering algorithm.

Input: $T_1 = (V, E)$, and G.
Output: Clusters C_g for $g = 1, 2, \ldots, G$.
1. Construct a positive $Q \times Q$ $\mathcal{W}_{i,j}$, where $\mathcal{W}_{i,j}$ is the similarity matrix between v_i and v_j.
2. Compute the diagonal matrix $\mathcal{D}$, whose elements are corresponding to $\mathfrak{d}_{ii} = \sum_j \mathsf{w}_{i,j}$.
3. Compute the normalized graph Laplacian matrix, as $\mathcal{L} = \mathcal{D}^{-0.5}\mathcal{W}\mathcal{D}^{-0.5}$.
4. Compute the top G eigenvectors of $\mathcal{L}$, where the obtained low-dimensional matrix is referred to as $\mathcal{F}$.
5. Considering the rows of $\mathcal{F}$ as CUEs, run the k-means algorithm to partition them into G clusters.

The optimal partition of P1 is based on minimizing the objective function Ncut, and this involves minimizing the edges' weights that need to be cut. However, it was proven in [39] that minimizing Ncut is NP-hard, and this means that finding the optimal solution for P1 is computationally prohibitive. Fortunately, the spectral clustering algorithm can provide a loose solution within polynomial time to optimize P1. Here, "loose" means relaxing the discrete optimization problem to the real number field and then using a heuristic approach to reconvert it into a discrete solution [40]. The optimization problem obtained after conducting Ncut relaxation [41] is formulated as

$$\min_{\mathcal{F} \in \mathbb{R}^{K \times G}} Tr(\mathcal{F}^t \mathcal{L} \mathcal{F}) \text{ Subject to } \mathcal{F}^t \mathcal{F} = \mathcal{I}, \tag{16}$$

where $\mathcal{L}$ is the normalized graph Laplacian matrix, with $\mathcal{L} = \mathcal{D}^{-0.5}\mathcal{W}\mathcal{D}^{-0.5}$; $\mathcal{W}$ is the weight matrix identified in P1; $\mathcal{I}$ is the identity matrix; $\mathcal{D}$ is a diagonal matrix whose elements are the degrees of the graph vertices, and it corresponds to $\mathfrak{d}_{ii} = \sum_{j=1}^{Q} \mathsf{w}_{i,j}$; $\mathcal{F}$ is the spectral embedding matrix; $Tr(.)$ denotes the trace of a matrix; and t denotes the matrix transposition operation. The algorithm in [42] is adopted to provide a suboptimal user clustering solution, and it consists of two main steps. First, embedding the data points (CUEs in our algorithm) are embedded in a low dimensional space by using the eigenvectors of a normalized graph Laplacian matrix. Second, a classical clustering algorithm such as K-means is applied. Spectral clustering can capture the manifold structure of data, which is impossible to accomplish by using only K-means clustering algorithms [43]. The graph-based CUE clustering algorithm is outlined in Table 1.

B. D2D PAIR CLUSTERING

1) GRAPH CONSTRUCTION

The D2D pair interference graph is constructed by the BS, and this stated as problem P2 as follows:

P2: A graph $T_2 = (V, E)$ is given, where $V = \mathsf{v}_1, \ldots, \mathsf{v}_Z$ each vertex from set V represents a D2D pair, and each edge from set E has a weight. An edge weight represents interference between two D2D pairs. The interference from v_i to v_j is denoted as $\mathsf{w}_{i,j}$, while the interference from v_j to v_i

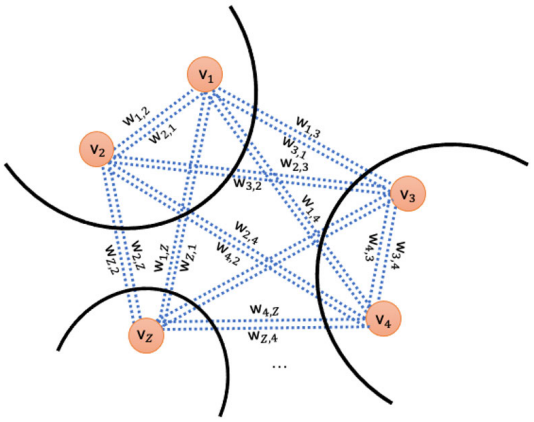

FIGURE 5. The interference graph of D2D pairs.

is denoted as $\mathsf{w}_{j,i}$, as shown in Fig. 5. When $\mathsf{w}_{i,j} = 0$ indicates that there is no interference from v_i to v_j.

The goal of P2 is to partition graph T_2 into G disjoint clusters $\{C_1, \ldots, C_G\}$, where the highly interfered D2D pairs are placed in different clusters to eliminate inter-cluster and intra-cluster interferences. To reach this goal, the sum of the weights of edges across different clusters, i.e., $\sum_{i\in C_i, j\in C_j, C_i\neq C_j} \mathsf{w}_{i,j} + \mathsf{w}_{j,i}$, is maximized. Here, the edge weight between v_i and v_j represents the interference from v_i to v_j, which is expressed as $H_{i,j}$ and defined as the same equation used in (4).

2) ALGORITHM DESIGNING

Problem P2 is equivalent to the MAX K-CUT problem in graph theory [44], [45], and it was proven in [44] that problem P2 is NP-hard, which means that finding the optimal solution for P2 is computationally prohibitive. Therefore, a simple heuristic algorithm [45] is adopted to cope with P2 and efficiently approximate the optimal solution. The algorithm idea is to iteratively assign D2D pairs into the clusters such that, at each step, the increased inter-cluster weight is computed. Then, D2D pair z is assigned to the cluster with minimum mutual interference $C_g^* = arg\min \sum_{z=1,\acute{z}\in C_g}^{Z} (\mathsf{w}_{z,\acute{z}} + \mathsf{w}_{\acute{z},z})$. The graph-based D2D pairs clustering algorithm is outlined in Table 2.

C. CLUSTERS MATCHING

1) GRAPH CONSTRUCTION

After performing CUE clustering and D2D pair clustering, one-to-one matching is applied to match each D2D pair cluster to its best CUE cluster for spectrum sharing under a single NOMA cluster. The cluster graph is stated as problem P3 as follows:

P3: A bipartite graph $T_3 = (V, E)$ is given, the vertex set $V = C_{D2D} \cup C_{CUE}$. $C_{D2D} = \{\mathsf{v}_{D2D}^1, \ldots, \mathsf{v}_{D2D}^G\}$ and $C_{CUE} = \{\mathsf{v}_{CUE}^1, \ldots, \mathsf{v}_{CUE}^G\}$ are disjoint sets, where $C_{D2D} \cap C_{CUE} = \phi$, v_{D2D} represents a cluster of D2D pairs, and v_{CUE} represents a cluster of CUEs, as shown in Fig. 6. A weighed edge

TABLE 2. Graph-based D2D pairs clustering algorithm.

Input: $T_2 = (V, E)$, and G.
Output: : C_g for $g = 1, 2, \ldots, G$.
1. Initialize $C_g = \Phi, \forall g = 1, 2, \ldots, G$.
2. Arbitrarily assign one D2D pair into each of the G cluster.
3. For D2D pair z is not in any cluster Do
4. For $g = 1 : G$ Do
5. Compute the interference using $\mathcal{W}_{z,\acute{z}} = \sum_{z=1\acute{z}\in C_g}^{Z} \mathsf{w}_{z,\acute{z}} + \mathsf{w}_{\acute{z},z}$.
6. End For
7. Assign the z-th D2D pair to C_g^*-th cluster with $C_g^* = arg\min \sum_{\acute{z}\in C_g} (\mathsf{w}_{z,\acute{z}} + \mathsf{w}_{\acute{z},z})$.
8. End For

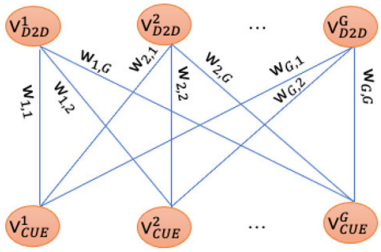

FIGURE 6. Bipartite graph for clusters matching.

represents the normalized channel gain from v_{D2D}^j to v_{CUE}^i and is denoted as $\mathsf{w}_{i,j}$.

The goal of P3 is to match each $\mathsf{v}_{D2D} \in C_{D2D}$ with one $\mathsf{v}_{CUE} \in C_{CUE}$, where the D2D pair cluster with its matched CUE cluster has high channel correlation while the D2D pair cluster with other CUE clusters has low channel correlation for mitigating the inter-cluster interference. To reach this goal, the sum of the weights of the edges between C_{D2D} and C_{CUE} i.e. $\sum_{i\in C_{D2D}, j\in C_{CUE}} \mathsf{w}_{i,j}$, is maximized.

The normalized channel correlation between a CUE cluster and a D2D cluster is identified as

$$\mathsf{w}_{i,j} = \frac{\mathsf{v}_{D2D}^i \mathsf{v}_{CUE}^j}{\mathsf{v}_{D2D}^i \sum_{\acute{j}=1,\acute{j}\neq j}^{G} \mathsf{v}_{CUE}^{\acute{j}}}, \tag{17}$$

where $\mathsf{v}_{D2D}^i \mathsf{v}_{CUE}^j$ represents the channel correlation between v_{D2D}^i and v_{CUE}^j; and $\mathsf{v}_{D2D}^j \sum_{\acute{j}=1,\acute{j}\neq j}^{G} \mathsf{v}_{CUE}^{\acute{j}}$ represents the channel correlation between v_{D2D}^i and all $\mathsf{v}_{CUE}^{\acute{j}}$.

2) ALGORITHM DESIGNING

The Hungarian algorithm [46] is an efficient way to solve problem P3 within polynomial time, and it can be used to find maximum-weight matches in the bipartite graph. This identifies the best set of users (i.e., CUEs and D2D pairs) for spectrum sharing.

D. COMPLEXITY ANALYSIS

Compared with exhaustive search, the time complexity of the proposed interference-aware graph-based clustering algorithm of is reduced. The proposed user clustering algorithm has a polynomial complexity in $O(Q^3 + (Z^2/2 + Z/2 + G) + G^3)$, where the first term corresponds to the complexity of the

spectral clustering used for CUE clustering [47], the second term corresponds to the heuristic algorithm used to solve the problem of multi-way MAX K-CUT [45], which is also used for D2D pair clustering, and the third term corresponds to the Hungarian algorithm used for clusters matching [48].

IV. NOMA-MIMO HB DESIGN

After performing the user clustering and before the signals are transmitted, the BS applies NOMA-MIMO HB which involves the design of a single beamforming vector to support multiple CUEs. The proposed design for the HB architecture is composed of a high-dimensional analog precoder and a low-dimensional digital precoder [19], where the standard two-step HB scheme [49] is considered, as illustrated in Fig. 2. The analog beamforming vector for each cluster is generated based on the channel gain of the strongest CUE within each cluster, which is known as the cluster-head Γ. For the digital precoding, ZF beamforming based on the equivalent channel gain is adopted to eliminate inter-beam interference. For analog precoding, phase shifters are employed to adjust the links of the N_{RF} chain with N_{BS} antennas.

In this paper, fully-connected and partially-connected HB architectures are considered. In fully-connected architecture, the phase shifter attaches each RF chain to all antennas and the number of phase shifters is equal to the number of antennas, as shown in Fig. 2. In partially-connected architecture, the phase shifter attaches each RF chain to a subset of antennas, and only M phase shifters are needed, where $M = \frac{N_{BS}}{N_{RF}}$. In general, a partially-connected HB architecture is simpler to implement and is expected to be more energy-efficient, although there could be some performance losses when compared to the fully-connected HB architecture [10].

A. ANALOG PRECODING

The $N_{BS} \times N_{RF}$ matrix is obtained by the analog precoder and is denoted as A. For fully-connected architecture,

$$A^{fully} = [\bar{a}_1, \bar{a}_2, \ldots, \bar{a}_{N_{RF}}], \tag{18}$$

where the elements of $\bar{a}_n \in N_{BS} \times 1$ for $n = 1, 2, \ldots, N_{RF}$ have the same amplitude $\frac{1}{\sqrt{N_{BS}}}$ but different phases [33]. For partially-connected,

$$A^{partially} = \begin{pmatrix} \bar{a}_1 & 0 & \cdots & 0 \\ 0 & \bar{a}_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \bar{a}_n \end{pmatrix}, \tag{19}$$

where the elements of $\bar{a}_n \in N_{BS} \times 1$ for $n = 1, 2, \ldots, N_{RF}$ have the same amplitude $\frac{1}{\sqrt{M}}$ but different phases [16]. In this paper, B-bit quantized phase shifters [50] is proposed for analog precoding. The phase set of B-bits composed of $\mathbb{A} = \{\frac{2\pi n}{2^B} : n = 0, 1, \ldots, 2^{B-1}\}$ and $\bar{a}_n$ elements of A belongs to the quantized phase $\frac{1}{\sqrt{N_{BS}}}\{e^{j\frac{2\pi n}{2^B}} : n = 0, 1, \ldots, 2^{B-1}\}$ in the case of a fully-connected architecture and $\frac{1}{\sqrt{M}}\{e^{j\frac{2\pi n}{2^B}} : n = 0, 1, \ldots, 2^{B-1}\}$ in the case of a partially-connected architecture [10]. Specifically, the analog precoding design is based on maximizing the array gain of Γ in the g-th cluster $H_{CUE(\Gamma,g)}$ for $g = 1, 2, \ldots, G$. For a fully-connected architecture [10], [51], the i-th $\bar{a}_n$ element of A, where $i = 1, 2, \ldots, N_{BS}$, can be expressed as

$$\bar{a}_{g(i)} = \frac{1}{\sqrt{N_{BS}}} e^{j\frac{2\pi \hat{n}}{2^B}}, \tag{20}$$

where

$$\hat{n} = arg\min |\angle H_{CUE(\Gamma,g)}(i) - \frac{2\pi n}{2^B}|. \tag{21}$$

The design for a fully-connected analog precoder is illustrated in Table 3. Similarly, for partially-connected architecture, the i-th $\bar{a}_n$ element of $\mathbb{A}$, where $i = (g-1)M+1, (g-1)M+2, \ldots, gM$ can be expressed as

$$\bar{a}_g(i) = \frac{1}{\sqrt{M}} e^{j\frac{2\pi \hat{n}}{2^B}}, \tag{22}$$

where $\hat{n}$ is defined using the same equation in (20).

TABLE 3. Analog precoder design algorithm.

Input: B, N_{BS}, N_{RF}, and $H_{CUE(\Gamma,g)}$ for $g = 1, 2, \ldots, G$.
Output: Analog precoder matrix A of size $N_{BS} \times N_{RF}$.

1. Initialize $A = zeros(N_{BS} \times N_{RF})$;
2. Define the phase set $\mathbb{A} = \frac{2\pi n}{2^B} : n = 0, 1, \ldots, 2^{B-1}$
3. For $g = 1$ to N_{RF} Do
4. Extract phases of $H_{CUE(\Gamma,g)}$, where $phase = \angle(H_{CUE(\Gamma,g)})$;
5. Set $QuantizedPhase = zeros(N_{BS})$;
6. For $n = 1$ to N_{BS} Do
7. $[\sim, i] = \min phase(n) - \mathbb{A}$
8. $QuantizedPhase(n) = \mathbb{A}(i)$
9. End For
10. $A(:, g) = e(j * QuantizedPhase)$, where j is an imaginary element.
11. End For

B. DIGITAL PRECODING

After analog precoding is performed, the equivalent channel matrix is obtained and can be expressed as $\bar{H} = \{\bar{H}_{CUE(\Gamma,1)}, \bar{H}_{CUE(\Gamma,2)}, \ldots, H_{CUE(\Gamma,G)}\}$ for $g = 1, 2, \ldots, G$, where the equivalent channel vector is expressed as $\bar{H}_{CUE(\Gamma,k)} = H_{CUE(\Gamma,k)}A$. Without loss of generality, fully-digital ZF beamforming is adopted to eliminate the inter-beam interference as a low-dimensional baseband digital precoder based on the strongest equivalent channel [51]. However, the digital precoding matrix of size $N_{RF} \times Q$ is generated by

$$\bar{D} = [\bar{d}_1, \bar{d}_2, \ldots, \bar{d}_G] = \bar{H}^H(\bar{H}\bar{H}^H)^{-1}. \tag{23}$$

The summarized digital precoder design is shown in Table 4.

TABLE 4. Digital precoder design algorithm.

Input: N_{BS}, N_{RF}, $\bar{H}$, and A.
Output: Digital precoder matrix D of size $N_{RF} \times Q$
1. Define $\bar{D} = \bar{H}^H(\bar{H}\bar{H}^H)^{-1}$;
2. Normalize $\bar{D} = (\bar{D}) * repmat(\sqrt{\sum(A * \bar{D})^2}, 1), N_{RF}, 1)$;
3. Initialize $d = zeros(N_{RF} * Q)$;
4. $D(:, cluster(:,1)) = \bar{D}$; where $cluster(:,1) = cluster_\Gamma$
5. For $g = 1$ to N_{RF} Do
6. $clusterg = cluster(g,:)$;
7. For $ng = 2$ to $length(clusterg)$ Do
8. $D(:, clusterg(ng) = D(:, cluster(g,1))$;
9. End For
10. End For

V. OPTIMIZED POWER ALLOCATION

At the receiver, to decode the superimposed signal and eliminate the intra-beam interference, the CUEs apply SIC. In the g-th beam, $CUE(g,k)$ can eliminate the intra-beam interference caused by the BS, which is the desired signal for the $CUE(g,\acute{k})$ if all $\acute{k}$-th channel gains $<$ the k-th channel gains [14], where CUEs with higher channel gains are assigned with lower power coefficients; then, each CUE decodes the signal by considering weaker signals in the X_g as interference, and are not decoded. The decoded signal can be either the desired signal or can be subtracted from the X_g. The decoding process continues until each CUE decodes its signal successfully [52]. On the other hand, CUEs with lower channel gains are assigned with higher power coefficients; then, they detect signals directly by treating the signals of CUEs are as noise [53]. In addition, to manage the intra-cluster interference caused by DUE_{Tx}s to CUEs in the g-th cluster, the transmission power of each DUE_{Tx} is kept under a predefined threshold th^g, as defined in C4 in (12).

A. POWER ALLOCATION PROBLEM FORMULATION

The power allocation problem in this scenario is transformed into the problem of maximizing the network spectral efficiency while guaranteeing the QoS requirements for both CUEs and D2D pairs. The sum-rate of the CUE and DUE maximization problem in (13) can be reformulated as follows:

$$\{P^*_{CUE}, P^*_{DUE}\} = arg\max(\sum_{g=1}^{G}(\sum_{k=1}^{K} R_{CUE(g,k)} + \sum_{m=1}^{M} R_{DUE(g,m)})), \quad (24)$$

Subject to $C1$, $C2$, $C3$, $C4$, $C5$, and $C6$. Since the problem in (24) is NP-hard, the PSO is adopted to solve this problem in polynomial time. PSO is a population-based intelligent stochastic algorithm for global optimization that was developed in 1995 [54]. Generally, PSO initializes a population of random particles within the search area. During every iteration, each particle adjusts its position according to the corresponding particle velocity. The new position is evaluated according to the fitness function, which is determined by the objective function. The particle's speed and position are affected by two factors. One is the optimal solution found by the current population which is denoted as "personal best" (*PBest*), and the other is the optimal solution found by all populations, which is denoted as "global best" (*GBest*). The update process is repeated iteratively until either an optimal *GBest* is obtained or a fixed number of iterations is completed.

B. OPTIMIZED POWER ALLOCATION BASED ON PSO

The proposed PSO-based power allocation algorithm solves the problem in (24), where the objective function in (24) is defined as the fitness function. The PSO algorithm is based on five main steps as follows:

1) GENERATION OF PARTICLE POSITION AND VELOCITY

Each particle is represented by its position and velocity during the t-th iteration. The particle position represents the power coefficient assigned to each of the Q CUEs and Z D2D pairs. The position of $particle(i)$ is represented by a vector $\mathbb{X}$ in $\mathbb{D}$-dimensional space as

$$\mathbb{X}_i(t) = [x_i^{CUE_1}, \ldots, x_i^{CUE_Q}, x_i^{DUE_1}, \ldots, x_i^{DUE_Z}]^T, \quad for\ i = 1, 2, \ldots, pop, \quad (25)$$

where pop is the number of particles; $\mathbb{D}$ represents the total number of power coefficients needed and is equal to $Q + Z$, and $x_i \in [P_{min}, P_{max}]$, which determines the power range allowed to be assigned to CUEs and D2D pairs. Constraints $C5$ and $C6$ in (24) are satisfied by the power range determined by P_{min} and P_{max}. This search space provides reasonable parameters that need to be optimized, and these in turn yields the optimized power coefficients. The velocity of $particle(i)$ is represented by a vector $\mathbb{V}$ in $\mathbb{D}$-dimensional space as where $v_i \in [\mathbb{V}_{min}, \mathbb{V}_{max}]$, which determines the velocity range of a particle. Since the search space is defined by $[P_{min}, P_{max}]$, then $\mathbb{V}_{max} = T(P_{max} - P_{min})$, where $0.1 \leq T \leq 1.0$, and $\mathbb{V}_{min} = -\mathbb{V}_{max}$.

$$\mathbb{V}_i(t) = [v_i^{CUE_1}, \ldots, v_i^{CUE_Q}, v_i^{DUE_1}, \ldots, v_i^{DUE_Z}]^T, \quad for\ i = 1, 2, \ldots, pop. \quad (26)$$

2) DETERMINING *GBest* AND *PBest*

Each particle position is evaluated by the fitness function. Here, the fitness function evaluates the sum-rate achieved with the power coefficients suggested by the $particle(i)$ positions, and the maximum sum-rate achieved by the current population is stored as $PBest_i$ and recorded as,

$$PBest_i(t) = [pb_i^{CUE_1}, \ldots, pb_i^{CUE_Q}, pb_i^{DUE_1}, \ldots, pb_i^{DUE_Z}]^T, \quad for\ i = 1, 2, \ldots, pop. \quad (27)$$

The maximum sum-rate achieved among all iterations is recorded as

$$GBest(t) = \max(PBest_i(1), PBest_i(2), \ldots, PBest_i(t)), \quad for\ t = 1, 2, \ldots, iter, \quad (28)$$

where $iter$ is the number of iterations.

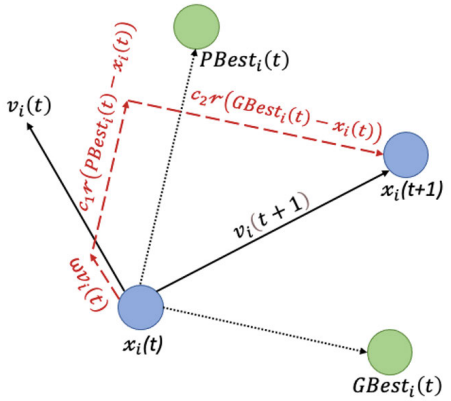

FIGURE 7. Illustration of the PSO position and velocity update processes.

3) UPDATING VELOCITY

The particle velocity is updated when entering a new iteration according to

$$v_i(t+1) = \omega v_i(t) + c_1 r(PBest_i(t) - x_i(t)) + c_2 r(GBest_i(t) - x_i(t)), \quad (29)$$

where ω is the inertia weight, r is a random number in the range [0, 1], c_1 is the self-confidence parameter, and c_2 is the swarm influence parameter.

Three components are affecting the particle velocity formula towards the optimal solution. The first component $\omega v_i(t)$ is used to balance deep search and breadth search. The second component $c_1 r(PBest_i(t) - x_i(t))$ is used to emphasize the capabilities of particles to search in the local area. The third component $c_2 r(GBes_i(t) - x_i(t))$ is used to emphasize the capabilities of particles to search in the global area, as shown in Fig. 7.

4) UPDATING PARTICLE POSITIONS

The position of each particle is updated using the new velocity vector for that particle according to

$$x_i(t+1) = x_i(t) + v_i(t+1). \quad (30)$$

An illustration of the PSO-based position and velocity update processes is shown in Fig. 7.

5) HANDLING CONSTRAINTS

Constraints are used to check the feasibility of the obtained particles positions. On this basis, the constraints $C1$, $C2$, and $C3$ in (24) are used to define the feasible search space so that only particles remaining in the feasible space are considered to determine the new values of *PBest* and *GBest*.

The flowchart of the proposed power allocation algorithm based on PSO is shown in Fig. 8.

VI. SIMULATION AND RESULTS

The proposed resource allocation algorithm for D2D communications at mmWave underlaying MIMO-NOMA cellular network is evaluated in terms of spectrum and energy efficiency. The common simulation parameters are illustrated in Table 5. In particular, the carrier frequency is 28 GHz, which is commonly used for mmWave broadband service. The bandwidth is normalized to 1 GHz, and this coincides with the data rate formulation in (10) and (11). The BS is equipped with a ULA of $N_{BS} = 64$ antennas. The DUE_{Tx} is equipped with 4 antennas. CUEs and D2D pairs are randomly distributed in a cell with a radius of 200 meters, and the maximum distance between D2D pairs is 30 meters. For $CUE(g, k)$ and $DUE(g, m)$, the channel vector is generated based on (2) and (4), respectively. Here, we assume that he number of scatters = 3, and thus, there is the number of paths for $L_{g,k} = 3$ and $L_{g,m} = 3$, including one LoS component for $l = 1$ and NLoS components for $2 < l < L$. The

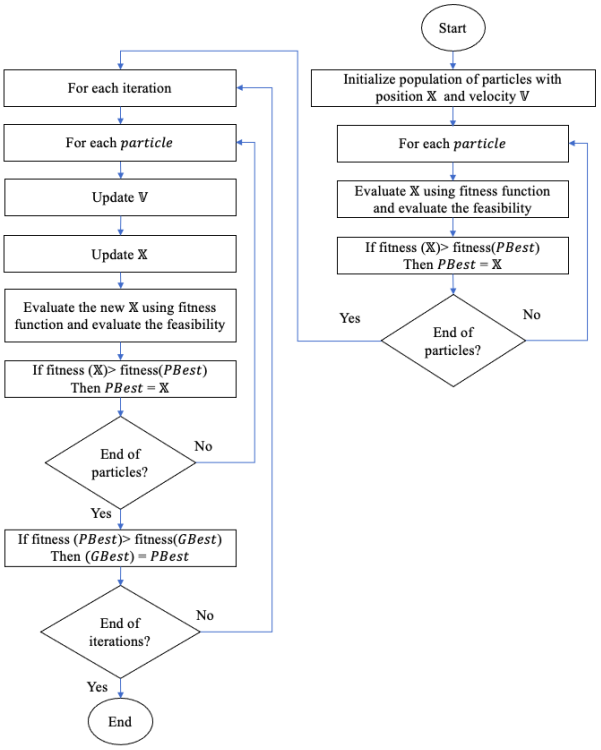

FIGURE 8. The flowchart of the PSO-based power allocation algorithm.

TABLE 5. Common simulation parameters.

Parameter	Value
Carrier frequency, f	28 GHz
Bandwidth	1 GHz
Small cell coverage	200 meters
The maximum distance between D2D pairs	30 meters
The number of BS antennas, N_{BS}	64
The number of DUE_{Tx} antennas, N_D	4
The number of RF china, N_{RF}	4 and 8
The phase shifters resolution, B	4 bit
Maximum transmit BS power, $P_{BS_{max}}$	35 dBm
Maximum transmit DUE_{Tx} power	15 dBm
PL exponent, $\delta \in \{\text{LOS, NLOS}\}$	2.1, 3.4
Population size, *pop*	50
The number of iterations, *iter*	100

path-loss exponent for LOS = 2.1 and NLOS = 3.4. For simplicity, we assume the minimal SINR required for each CUE and D2D pair to guarantee the QoS is $\gamma_{CUE(g,k),min} = \gamma_{DUE(g,m),min} = \gamma_{ZF}/10$, where γ_{ZF} is the minimal SINR among all CUEs in fully-digital ZF beamforming. We assume that the predefined interference threshold $th^g = 10$ dBm. The spectrum efficiency (bit/s/Hz) is defined as the total achievable sum-rate R in (12). The energy efficiency in (bit/s/dBm) is determined as the ratio between the total sum-rate and the total consumed power [7], and it is formulated as

$$Energy\ efficiency = \frac{R}{P_{Total}}, \tag{31}$$

$$P_{Total} \quad P_{BS} + P_{D2D} + P_{RF} + P_{PS} + P_{BB}, \tag{32}$$

where P_{D2D} is the total consumed power by admitted DUE_{Tx}, P_{RF} is the total power used by RF chains, P_{PS} is the total power used by phase shifters, and P_{BB} is power used by the baseband. Specifically, the values in [55] are adopted, where $P_{RF} = 25$ dBm, $P_{PS} = 16$ dBm, and $P_{BB} = 23$ dBm.

In this paper, five schemes are considered for the comparison as follows: 1) "D2D communications underlaying fully-digital ZF beamforming MIMO", where one dedicated RF chain is required to each BS antenna and each D2D pair is one-to-one matched with one CUE based on their channel correlations; 2) "D2D communications underlaying fully-connected HB MIMO-NOMA", where the proposed interference-aware graph-based user clustering and optimized power allocation based on PSO are applied; 3) "D2D communications underlay partially-connected HB MIMO-NOMA"; 4) "D2D communications underlaying fully-connected HB MIMO-OMA"; and 5) "D2D communications underlaying partially-connected HB MIMO-OMA", where the system model with MIMO-OMA scheme is similar to "MIMO-NOMA", but OMA is performed for CUEs in each beam and TDMA is adopted as OMA scheme. Specifically, the time slot in each beam is equally divided among Q CUEs, and each D2D pair is one-to-one matched to each CUE based on their channel correlations. In the case of more number of D2D pairs than the number of CUEs, virtual CUEs are presumed.

Fig. 9 and Fig. 10 show the convergence of the proposed optimized power allocation based on PSO for the D2D communications at mmWave underlaying MIMO-NOMA cellular networks, where $Q = 8$, $Z = 20$, $N_{RF} = 4$, and signal-to-noise-ratio (SNR) = 0, for fully-connect and partially-connected HBs, respectively. The spectral efficiency tends to be stable after 12 iterations, and this confirms the convergence of the proposed power allocation based on PSO.

Fig. 11 shows the spectrum efficiency of the five schemes versus the SNRs, where $Q = 8$, $Z = 20$, and $N_{RF} = 4$. The proposed resource allocation algorithm for D2D communications at mmWave underlaying MIMO-NOMA under fully-connected HB cellular networks achieves greater spectrum efficiency compared to D2D communications at mmWave underlaying MIMO-OMA cellular networks under fully-connected HB. Furthermore, the proposed resource allocation

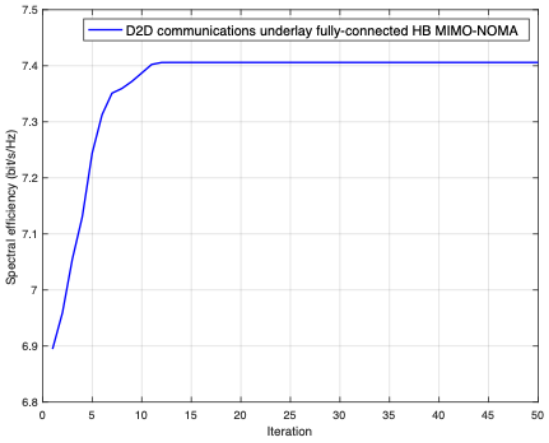

FIGURE 9. The convergence of the optimized power allocation for D2D communication underlaying fully-connected MIMO-NOMA HB.

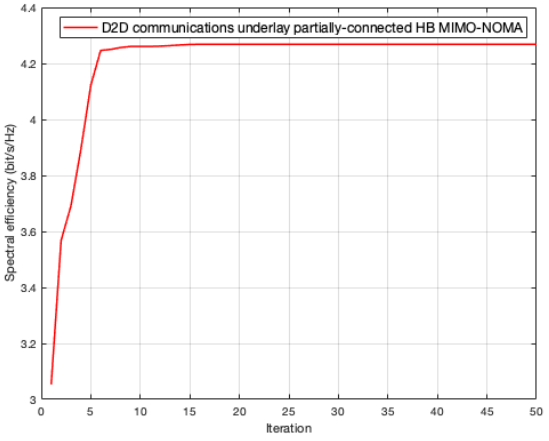

FIGURE 10. The covergence of the optimized power allocation for D2D communications underlaying partially-connected MIMO-NOMA HB.

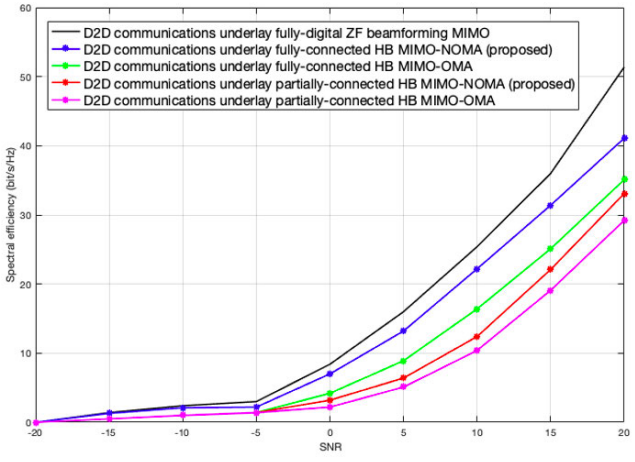

FIGURE 11. Spectral efficiency versus SNRs.

algorithm for D2D communication at mmWave underlaying MIMO-NOMA under partially-connected HB cellular networks achieves greater spectrum efficiency compared to

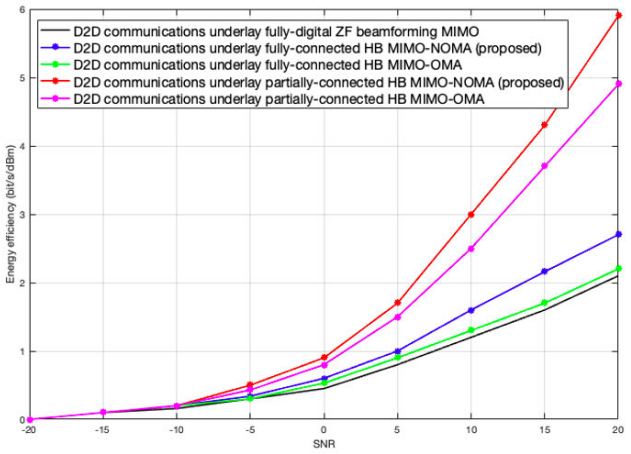

FIGURE 12. Energy efficiency versus SNRs.

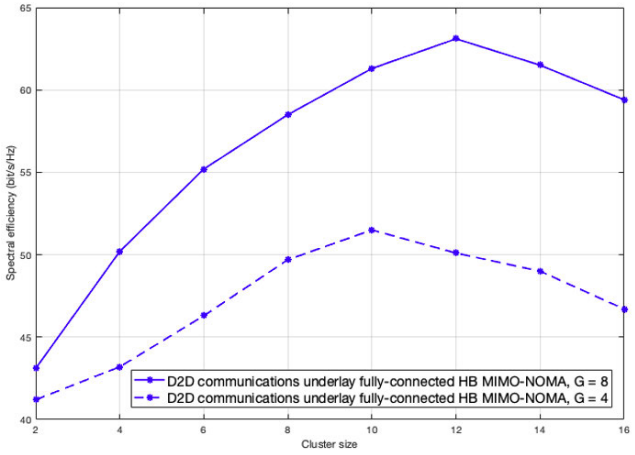

FIGURE 13. Spectral efficiency versus CUE cluster sizes.

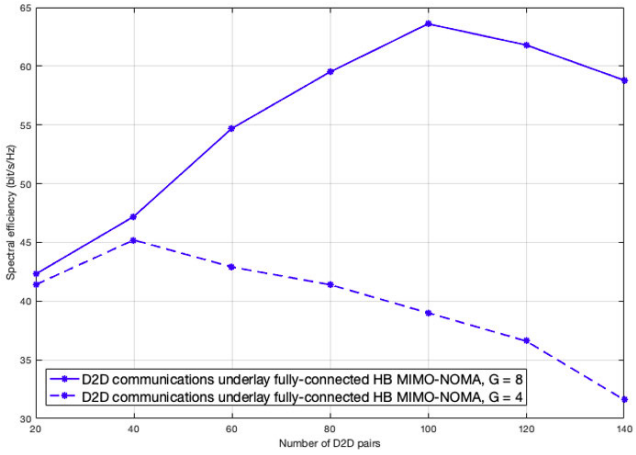

FIGURE 14. Spectral efficiency versus the number of D2D pairs.

D2D communications at mmWave underlaying MIMO-OMA cellular networks under partially-connected HB. Generally, D2D communications with MIMO-NOMA HB architecture achieve greater spectral efficiency than D2D communications with MIMO-OMA HB. Furthermore, the fully-connected MIMO-NOMA HB achieves greater spectrum efficiency than the partially-connected MIMO-NOMA HB, since the RF chains in the fully-connected leverage the full antenna array multiplexing gains. In addition, the D2D communications under fully-digital ZF beamforming MIMO achieve the best spectrum efficiency compared to other schemes since a dedicated RF chain is used to serve each CUE.

Fig. 12 shows the energy efficiency of the five schemes versus the SNRs, where $Q = 8$, $Z = 20$, and $N_{RF} = 4$. The proposed resource allocation algorithm for D2D communication at mmWave underlaying MIMO-NOMA cellular networks achieves the best energy efficiency. It is worth noting that although the D2D communication underlaying fully-digital ZF beamforming obtains the best spectral efficiency performance compared to other schemes, it has the least energy efficiency performance. This happens because the number of RF chains is equal to the number of BS antennas in fully-digital MIMO scheme, which results in very high energy consumption, e.g., 25 dBm for each RF chain. While in MIMO-NOMA HB, the number of RF chains is much less than the number of BS antennas. Therefore, it is possible to greatly reduce the energy consumption caused by RF chains in MIMO-NOMA HB. In addition, the partially-connected HB achieves greater energy efficiency than the fully-connected HB, since the partially-connected HB adopts a smaller number of phase shifters.

Fig. 13 shows the effect of the cluster size which is represented as the number of CUEs per cluster on the spectral efficiency, where $Z = 20$ and $SNR = 20$. For comparison purposes, the number of CUEs in different clusters is equal. The two scenarios are considered when $N_{RF} = G = 4$ and 8. The spectral efficiency for the proposed resource allocation for D2D communications at mmWave underlaying MIMO-NOMA cellular network increases almost linearly with the cluster size until approaching a particular cluster size, 10 and 12 at the number of clusters 4 and 8, respectively. After that, spectral efficiency starts to decrease due to the growth of interference signals. Here, when the number of CUEs is increased dramatically, the lower is the spectral efficiency, which implies a tradeoff between spectral efficiency and the number of admitted CUEs. The better spectral efficiency obtained with large number of clusters with less CUEs within. Therefore, it is crucial to select the exact cluster size.

Fig. 14 shows the effect of different numbers of D2D pairs on the spectral efficiency, where $Q = 20$ and $SNR = 20$. The two scenarios are considered when $N_{RF} = G = 4$ and when 8. The proposed D2D communications at mmWave underlaying MIMO-NOMA cellular networks achieves greater spectrum efficiency as the number of D2D pairs increases until approaching a particular number of D2D pairs, 40 and 100 at the number of clusters 4 and 8, respectively. After that, spectral efficiency starts to decrease due to the growth of

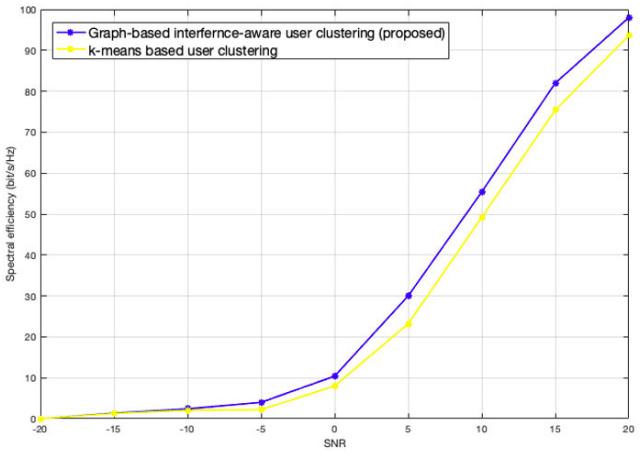

FIGURE 15. Spectral efficiency performance for graph-based interference-aware user clustering and k-means based user clustering.

interference signals. When the number of D2D pairs is high and the number of clusters is small, there is a large chance of decreasing the spectral efficiency while increasing the number of clusters helps to reduce the inference and increase the spectral efficiency. Therefore, it is crucial to select the exact number of D2D pairs that obtain the highest spectral efficiency.

To evaluate the user clustering algorithm in terms of spectral efficiency, a k-means based algorithm is considered for the comparison purposes. In this algorithm, the k-means clustering algorithm is used for CUE clustering and D2D pair clustering, and then Hungarian algorithm is used for one-to-one matching. Fig. 15 shows the spectral efficiency performance graph-based interference-aware user clustering algorithm and the K-means clustering for D2D communications underlaying MIMO-NOMA cellular network, where $G = 8$, $Q = 80$, $Z = 80$, and $SNR = 20$. The spectral efficiency of the proposed graph-based interference-aware user clustering exceeds K-means clustering for D2D communications at mmWave underlaying MIMO-NOMA cellular network.

VII. CONCLUSION AND FUTURE WORK

In this paper, we considered the integration of D2D communications at mmWave underlaying MIMO-NOMA cellular network to increase its spectral efficiency. To solve the NP-hard resource allocation problem, user grouping, beamforming, and power allocation are carefully designed in a tractable way to obtain a suboptimal solution. First, we proposed a user clustering algorithm based on graph theory that defines the best cluster of CUEs for MIMO-NOMA HB and the best user cluster (i.e., CUEs and DUEs) for spectrum sharing with eliminated inter-cluster interference caused by DUE_{Tx}s to CUEs and by DUE_{Tx}s to DUE_{Rx}s, as well as intra-cluster interferences caused by DUE_{Tx}s to DUE_{Rx}s. After that, we designed a MIMO-NOMA HB to transmit superimposed signals through beams to CUE clusters, where analog precoding is designed based on the cluster-head for each beam and the digital precoding is designed to eliminate the inter-beam interference by adopting ZF beamforming based on the equivalent channel gain in each beam. Next, SIC technology is adopted in NOMA for superimposed single decoding at CUEs and to eliminate the intra-beam interference. Finally, an optimized power allocation based on PSO is proposed for both CUEs and DUEs with the objectives of maximizing the spectral efficiency, while protecting CUEs from intra-cluster interference caused by DUE_{Tx} and guaranteeing QoS for CUEs and D2D pairs.

Simulation results demonstrate that the proposed resource allocation algorithm for D2D communications at mmWave underlaying MIMO-NOMA cellular network delivers greater spectral efficiency and energy efficiency than the conventional D2D communications that operate underlay MIMO-OMA cellular networks. In addition, simulation results display the effects of different cluster sizes and numbers of D2D pairs in a cluster on the spectral efficiency. It has shown that the proposed system model can utilize the spectrum to support large-scale users with improved spectral efficiency and energy efficiency. The results obtained in this paper can also be useful for the design of the future 5G cellular networks and pave the way for D2D communications to be implemented into the 5G cellular network.

In future work, the efficiency of the proposed algorithms will be studied under partial CSI knowledge instead of the full CSI knowledge we assumed in the proposed algorithm. Furthermore, research will be conducted on the optimized cluster size and the number of D2D pairs in each cluster. The system model of D2D communications underlaying MIMO-NOMA at mmWave in the uplink period is a challenging task and it will be investigated. In addition, we will consider the results obtained by the current simulation work as a dataset for a deep learning resource allocation approach for D2D communications at mmWave underlaying MIMO-NOMA cellular networks.

REFERENCES

[1] G. Forecast, "Cisco visual networking index: Global mobile data traffic forecast update 2017–2022," Cisco, San Jose, CA, USA, Tech. Rep. c11-738429, Feb. 2019, p. 2022.
[2] J. Li, X. Li, A. Wang, and N. Ye, "Performance analysis for downlink MIMO-NOMA in millimeter wave cellular network with D2D communications," *Wireless Commun. Mobile Comput.*, vol. 2019, pp. 1–11, Jun. 2019.
[3] R. W. Heath, Jr., N. Gonzalez-Prelcic, S. Rangan, W. Roh, and A. M. Sayeed, "An overview of signal processing techniques for millimeter wave MIMO systems," *IEEE J. Sel. Topics Signal Process.*, vol. 10, no. 3, pp. 436–453, Apr. 2016.
[4] M. Olyaee, M. Eslami, and J. Haghighat, "Performance of maximum ratio combining of fluctuating two-ray (FTR) mmWave channels for 5G and beyond communications," *Trans. Emerg. Telecommun. Technol.*, vol. 30, no. 10, p. e3601, Oct. 2019.
[5] J. G. Andrews, S. Buzzi, W. Choi, S. V. Hanly, A. Lozano, A. C. K. Soong, and J. C. Zhang, "What will 5G be?" *IEEE J. Sel. Areas Commun.*, vol. 32, no. 6, pp. 1065–1082, Jun. 2014.
[6] S. Kutty and D. Sen, "Beamforming for millimeter wave communications: An inclusive survey," *IEEE Commun. Surveys Tuts.*, vol. 18, no. 2, pp. 949–973, 2nd Quart., 2016.
[7] T. E. Bogale, X. Wang, and L. B. Le, "mmWave communication enabling techniques for 5G wireless systems: A link level perspective," in *MmWave Massive MIMO*. Amsterdam, The Netherlands: Elsevier, 2017, pp. 195–225.

[8] A. N. Uwaechia and N. M. Mahyuddin, "A comprehensive survey on millimeter wave communications for fifth-generation wireless networks: Feasibility and challenges," *IEEE Access*, vol. 8, pp. 62367–62414, 2020.
[9] Z. Zheng and H. Gharavi, "Spectral and energy efficiencies of millimeter wave MIMO with configurable hybrid precoding," *IEEE Trans. Veh. Technol.*, vol. 68, no. 6, pp. 5732–5746, Jun. 2019.
[10] L. Dai, B. Wang, M. Peng, and S. Chen, "Hybrid precoding-based millimeter-wave massive MIMO-NOMA with simultaneous wireless information and power transfer," *IEEE J. Sel. Areas Commun.*, vol. 37, no. 1, pp. 131–141, Jan. 2019.
[11] C. Hansen, "WiGiG: Multi-gigabit wireless communications in the 60 GHz band," *IEEE Wireless Commun.*, vol. 18, no. 6, pp. 6–7, Dec. 2011.
[12] L. Dai, B. Wang, Z. Ding, Z. Wang, S. Chen, and L. Hanzo, "A survey of non-orthogonal multiple access for 5G," *IEEE Commun. Surveys Tuts.*, vol. 20, no. 3, pp. 2294–2323, 3rd Quart., 2018.
[13] Z. Wei, L. Yang, D. W. K. Ng, J. Yuan, and L. Hanzo, "On the performance gain of NOMA over OMA in uplink communication systems," *IEEE Trans. Commun.*, vol. 68, no. 1, pp. 536–568, Jan. 2020.
[14] L. Dai, B. Wang, Y. Yuan, S. Han, I. Chih-Lin, and Z. Wang, "Non-orthogonal multiple access for 5G: Solutions, challenges, opportunities, and future research trends," *IEEE Commun. Mag.*, vol. 53, no. 9, pp. 74–81, Sep. 2015.
[15] Y. Saito, Y. Kishiyama, A. Benjebbour, T. Nakamura, A. Li, and K. Higuchi, "Non-orthogonal multiple access (NOMA) for cellular future radio access," in *Proc. IEEE 77th Veh. Technol. Conf. (VTC Spring)*, Jun. 2013, pp. 1–5.
[16] X. Gao, L. Dai, S. Han, C.-L. I, and R. W. Heath, Jr., "Energy-efficient hybrid analog and digital precoding for mmWave MIMO systems with large antenna arrays," *IEEE J. Sel. Areas Commun.*, vol. 34, no. 4, pp. 998–1009, Apr. 2016.
[17] X. Wu, S. Tavildar, S. Shakkottai, T. Richardson, J. Li, R. Laroia, and A. Jovicic, "FlashLinQ: A synchronous distributed scheduler for peer-to-peer ad hoc networks," *IEEE/ACM Trans. Netw.*, vol. 21, no. 4, pp. 1215–1228, Aug. 2013.
[18] U. N. Kar and D. K. Sanyal, "An overview of device-to-device communication in cellular networks," *ICT Exp.*, vol. 4, no. 4, pp. 203–208, Dec. 2018.
[19] F. O. Ombongi, H. O. Absaloms, and P. L. Kibet, "Resource allocation in millimeter-wave device-to-device networks," *Mobile Inf. Syst.*, vol. 2019, pp. 1–16, Dec. 2019.
[20] S. Umrao, A. Roy, and N. Saxena, "Device-to-device communication from control and frequency perspective: A composite review," *IETE Tech. Rev.*, vol. 34, no. 3, pp. 286–297, May 2017.
[21] S. Sobhi-Givi, M. G. Shayesteh, and H. Kalbkhani, "Energy-efficient power allocation and user selection for mmWave-NOMA transmission in M2M communications underlaying cellular heterogeneous networks," *IEEE Trans. Veh. Technol.*, vol. 69, no. 9, pp. 9866–9881, Sep. 2020.
[22] J. Ning, L. Feng, F. Zhou, M. Yin, P. Yu, W. Li, and X. Qiu, "Interference control based on stackelberg game for D2D underlaying 5G mmWave small cell networks," in *Proc. IEEE Int. Conf. Commun. (ICC)*, May 2019, pp. 1–6.
[23] H. Sun, Y. Xu, and R. Q. Hu, "A NOMA and MU-MIMO supported cellular network with underlaid D2D communications," in *Proc. IEEE 83rd Veh. Technol. Conf. (VTC Spring)*, May 2016, pp. 1–5.
[24] S. M. A. Kazmi, N. H. Tran, T. M. Ho, A. Manzoor, D. Niyato, and C. S. Hong, "Coordinated device-to-device communication with non-orthogonal multiple access in future wireless cellular networks," *IEEE Access*, vol. 6, pp. 39860–39875, 2018.
[25] K. M. S. Huq, S. Mumtaz, J. Rodriguez, P. Marques, B. Okyere, and V. Frascolla, "Enhanced C-RAN using D2D network," *IEEE Commun. Mag.*, vol. 55, no. 3, pp. 100–107, Mar. 2017.
[26] L. Zhu, J. Zhang, Z. Xiao, X. Cao, D. O. Wu, and X.-G. Xia, "Millimeter-wave NOMA with user grouping, power allocation and hybrid beamforming," *IEEE Trans. Wireless Commun.*, vol. 18, no. 11, pp. 5065–5079, Nov. 2019.
[27] Z. Wei, L. Zhao, J. Guo, D. W. K. Ng, and J. Yuan, "Multi-beam NOMA for hybrid mmWave systems," *IEEE Trans. Commun.*, vol. 67, no. 2, pp. 1705–1719, Feb. 2019.
[28] K. Chandra, A. S. Marcano, S. Mumtaz, R. V. Prasad, and H. L. Christiansen, "Unveiling capacity gains in ultradense networks: Using mm-wave NOMA," *IEEE Veh. Technol. Mag.*, vol. 13, no. 2, pp. 75–83, Jun. 2018.
[29] Z. Wei, D. W. K. Ng, and J. Yuan, "NOMA for hybrid mmWave communication systems with beamwidth control," *IEEE J. Sel. Topics Signal Process.*, vol. 13, no. 3, pp. 567–583, Jun. 2019.
[30] Z. Xiao, L. Zhu, J. Choi, P. Xia, and X.-G. Xia, "Joint power allocation and beamforming for non-orthogonal multiple access (NOMA) in 5G millimeter wave communications," *IEEE Trans. Wireless Commun.*, vol. 17, no. 5, pp. 2961–2974, May 2018.
[31] L. Zhu, J. Zhang, Z. Xiao, X. Cao, and D. O. Wu, "Optimal user pairing for downlink non-orthogonal multiple access (NOMA)," *IEEE Wireless Commun. Lett.*, vol. 8, no. 2, pp. 328–331, Apr. 2019.
[32] W. Hao, M. Zeng, Z. Chu, and S. Yang, "Energy-efficient power allocation in millimeter wave massive MIMO with non-orthogonal multiple access," *IEEE Wireless Commun. Lett.*, vol. 6, no. 6, pp. 782–785, Dec. 2017.
[33] O. E. Ayach, S. Rajagopal, S. Abu-Surra, Z. Pi, and R. W. Heath, Jr., "Spatially sparse precoding in millimeter wave MIMO systems," *IEEE Trans. Wireless Commun.*, vol. 13, no. 3, pp. 1499–1513, Mar. 2014.
[34] S. Sun, G. R. MacCartney, and T. S. Rappaport, "Millimeter-wave distance-dependent large-scale propagation measurements and path loss models for outdoor and indoor 5G systems," in *Proc. 10th Eur. Conf. Antennas Propag. (EuCAP)*, Apr. 2016, pp. 1–5.
[35] M. 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, 2017.
[36] 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.
[37] R. Zhang, X. Cheng, Q. Yao, C.-X. Wang, Y. Yang, and B. Jiao, "Interference graph-based resource-sharing schemes for vehicular networks," *IEEE Trans. Veh. Technol.*, vol. 62, no. 8, pp. 4028–4039, Oct. 2013.
[38] K. Allab, L. Labiod, and M. Nadif, "Simultaneous spectral data embedding and clustering," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 29, no. 12, pp. 6396–6401, Dec. 2018.
[39] J. Shi and J. Malik, "Normalized cuts and image segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 22, no. 8, pp. 888–905, Aug. 2000.
[40] H. Jia, S. Ding, X. Xu, and R. Nie, "The latest research progress on spectral clustering," *Neural Comput. Appl.*, vol. 24, no. 7, pp. 1477–1486, 2014.
[41] E. P. Xing and M. I. Jordan, "On semidefinite relaxations for normalized k-cut and connections to spectral clustering," Univ. California Berkeley, Berkeley, CA, USA, Tech. Rep. UCB/CSD-03-1265, 2003.
[42] A. Y. Ng, M. I. Jordan, and Y. Weiss, "On spectral clustering: Analysis and an algorithm," in *Proc. Adv. Neural Inf. Process. Syst.*, 2002, pp. 849–856.
[43] S. Yoo, H. Huang, and S. P. Kasiviswanathan, "Streaming spectral clustering," in *Proc. IEEE 32nd Int. Conf. Data Eng. (ICDE)*, May 2016, pp. 637–648.
[44] S. Sahni and T. Gonzalez, "P-complete approximation problems," *J. ACM*, vol. 23, no. 3, pp. 555–565, Jul. 1976.
[45] R. Y. Chang, Z. Tao, J. Zhang, and C.-C.-J. Kuo, "Multicell OFDMA downlink resource allocation using a graphic framework," *IEEE Trans. Veh. Technol.*, vol. 58, no. 7, pp. 3494–3507, Sep. 2009.
[46] J. Munkres, "Algorithms for the assignment and transportation problems," *J. Soc. Ind. Appl. Math.*, vol. 5, no. 1, pp. 32–38, 1957.
[47] X. Chen, W. Hong, F. Nie, D. He, M. Yang, and J. Z. Huang, "Spectral clustering of large-scale data by directly solving normalized cut," in *Proc. 24th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, Jul. 2018, pp. 1206–1215.
[48] F. Sun, V. O. K. Li, and Z. Diao, "Modified bipartite matching for multiobjective optimization: Application to antenna assignments in MIMO systems," *IEEE Trans. Wireless Commun.*, vol. 8, no. 3, pp. 1349–1355, Mar. 2009.
[49] A. Alkhateeb, G. Leus, and R. W. Heath, Jr., "Limited feedback hybrid precoding for multi-user millimeter wave systems," *IEEE Trans. Wireless Commun.*, vol. 14, no. 11, pp. 6481–6494, Nov. 2015.
[50] X. Zhu, Z. Wang, L. Dai, and Q. Wang, "Adaptive hybrid precoding for multiuser massive MIMO," *IEEE Commun. Lett.*, vol. 20, no. 4, pp. 776–779, Apr. 2016.
[51] A. N. Uwaechia and N. M. Mahyuddin, "Spectrum and energy efficiency optimization for hybrid precoding-based SWIPT-enabled mmWave mMIMO-NOMA systems," *IEEE Access*, vol. 8, pp. 139994–140007, 2020.
[52] Y. Saito, A. Benjebbour, Y. Kishiyama, and T. Nakamura, "System-level performance evaluation of downlink non-orthogonal multiple access (NOMA)," in *Proc. IEEE 24th Annu. Int. Symp. Pers., Indoor, Mobile Radio Commun. (PIMRC)*, Sep. 2013, pp. 611–615.
[53] B. Di, L. Song, and Y. Li, "Sub-channel assignment, power allocation, and user scheduling for non-orthogonal multiple access networks," *IEEE Trans. Wireless Commun.*, vol. 15, no. 11, pp. 7686–7698, Nov. 2016.

[54] R. Eberhart and J. Kennedy, "A new optimizer using particle swarm theory," in *Proc. 6th Int. Symp. Micro Mach. Human Sci. (MHS)*, 1995, pp. 39–43.
[55] X. Gao, L. Dai, Y. Sun, S. Han, and I. Chih-Lin, "Machine learning inspired energy-efficient hybrid precoding for mmWave massive MIMO systems," in *Proc. IEEE Int. Conf. Commun. (ICC)*, May 2017, pp. 1–6.

SUHARE SOLAIMAN received the bachelor's degree in computer sciences from Taif University, Taif, Saudi Arabia, in 2007, and the master's degree in computer sciences from King Abdulaziz University, Jeddah, Saudi Arabia, in 2016, where she is currently pursuing the Ph.D. degree in computer science with the Department of Computer Science, Faculty of Computing and Information Technology. From 2008 to 2015, she was a Research Assistant with the Department of Computer Sciences, College of Computers and Information Technology, Taif University, where she is also working as a Lecturer. Her research interests include 5G cellular networks, NOMA, MIMO, mmWave channel modeling, the Internet of Things, and D2D communications.

LAILA NASSEF received the B.Sc. degree in electronics and communications and the M.Sc. degree in electronics and computers from Ain Shams University, Cairo, Egypt, in 1984 and 1989, respectively, and the Ph.D. degree in computer engineering from Anglia Polytechnic University, U.K., in 1994. She joined the National Telecommunications Institute, Cairo, in 1985. She joined the Institute of Statistical Studies and Research, Cairo University, Cairo, in 2001. She is currently an Associate Professor with the Department of Computer Science, Faculty of Computing and Information Technology, King Abdulaziz University, Jeddah, Saudi Arabia. Her research interests include wireless networks, cognitive radio networks, heterogeneous networks, smart grid communications, multiple access technologies, NOMA, MIMO, mmWave channel modeling, the Internet of Things, and Internet of Nano things.

ETIMAD FADEL (Member, IEEE) received the bachelor's degree in computer sciences from King Abdulaziz University, Jeddah, Saudi Arabia, with the Senior Project title ATARES: Arabic Character Analysis and Recognition in 1994, and the Integrated M.Phil./Ph.D. degree in computer science from De Montfort University (DMU), Leicester, U.K., in 2007. After getting promoted to an Assistant Professor, she was appointed as the Vice-Dean of the girl's section of FCIT from 2008 to 2010. She is currently working as an Associate Professor with the Computer Science Department, King Abdulaziz University. Her research interests include distributed systems, which are developed based on middleware technology. She is also looking into and working on wireless networks, the Internet of Things, and Internet of Nano-things. In addition, she is working on smart grids and HetNets.