IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received December 5, 2017, accepted January 16, 2018, date of publication February 1, 2018, date of current version March 28, 2018.

Digital Object Identifier 10.1109/ACCESS.2018.2799959

Impact of NOMA on Network Capacity Dimensioning for 5G HetNets

ANDREA S. MARCANO, (Member, IEEE), AND HENRIK L. CHRISTIANSEN, (Member, IEEE)
Department DTU Fotonik, Technical University of Denmark, 2800 Kongens Lyngby, Denmark
Corresponding author: Andrea S. Marcano (anmarc@fotonik.dtu.dk)

ABSTRACT Non-orthogonal multiple access (NOMA) has emerged as a key technology for boosting the capacity of 5G networks. Since, the latter are expected to be heterogeneous networks (HetNets), the performance of NOMA on 5G HetNets is highly anticipated. In this paper, we present a system-level analysis, focused on the capacity dimensioning, of a 5G HetNet with hybrid multiple access where NOMA and orthogonal multiple access coexist. We use dynamic power allocation and consider four generic pairing methods for NOMA: Hungarian, Gale–Shapley, random, and exhaustive. Through our results, we show that the optimal or close-to-optimal pairing methods offer the highest capacity gain (22%–24%) when the network cells are equally loaded. On the contrary, if the load is unequal and load balancing techniques are used, simpler pairing methods offer higher gains (approximately 29%). This leads to the idea of a flexible choice of the pairing method to be used for NOMA depending on the network load, thus achieving a balance between the network capacity gain and the complexity of the pairing method. In our network, for 100 cells, the combination of the Hungarian and the random method allows supporting 4% higher network traffic volume than if either of these two methods is exclusively used. Such gain can be translated into fewer cells needed for the same traffic volume, or higher traffic volume with the same number of cells. Furthermore, our results on network user dimensioning show that NOMA and HetNets can have the capacity to cope with the high data demand expected for 5G.

INDEX TERMS NOMA, capacity dimensioning, pairing methods, 5G, HetNets, hybrid multiple access.

I. INTRODUCTION

As the demand for digital content and services over the mobile networks continues to rise, same as the number of users/devices coming online, the current fourth generation (4G) of mobile networks are about to reach their capacity limit. It is expected that by 2022, there will around 9 billion mobile subscriptions, 8.3 billion mobile broadband subscriptions, and 6.2 billion unique mobile subscribers. With this high number of subscriptions, the mobile data traffic is also expected to grow, reaching values of 71 ExaBytes per month. Most of this high data volume is fueled by the video applications [1].

Therefore, one of the main improvements that comes with the deployment of the fifth-generation of mobile networks (5G), is higher capacity in comparison to 4G. 5G networks are expected to handle traffic 1000 times higher than 4G, and with connections up to 10 to 20 times faster. Moreover, near-zero latency, and energy saving and cost reduction are also requirements. With this, 5G welcomes the rapid development of the Internet of Things (IoT), connected homes, smart cities, autonomous driving, ultra-high definition (UHD) video and virtual reality, among others.

Two of the main solutions that have emerged as capacity boosters for 5G are non-orthogonal multiple access (NOMA) and massive deployment of small cells. With the integration of NOMA, the spectral efficiency can be increased, offering higher capacity than orthogonal multiple access (OMA) without increasing the available resources. This is possible through user pairing and multiplexing in the same time/frequency resources. However, in crowded scenarios where the network performance can significantly degrade, additional solutions might be needed to cope with the high traffic volume. For this, cell densification is a straightforward solution. By massively deploying small cells and tightly integrating them with the already deployed macro cells, the network load can be spread and enhance the quality of service (QoS).

The implementation of these solutions, lead to what is known as heterogeneous networks (HetNet) with hybrid multiple access (MA). That is, networks where macro and

small cells coexist, as well as different multiple access (MA) schemes.

A. MOTIVATION AND CONTRIBUTION

The implementation of NOMA in HetNets can offer extended benefits since both technologies share the objective of improving the spectral efficiency. The work in [2] proposes a resource management design for NOMA in HetNets and shows that NOMA based HetNets allow achieving a significantly higher performance than OMA based HetNets.

The benefits of NOMA and small cells for 5G have been studied separately in numerous research works.; [3]–[7] are some examples of such works. However, since the integration of small cells and NOMA is a natural one, research on how their combination influences the performance of a HetNet from a system-level perspective is needed. Moreover, research on how such combination influences the deployment and dimensioning of 5G HetNets is highly anticipated.

The purpose of this paper is to provide a system-level evaluation of the impact of NOMA in the downlink capacity dimensioning of a HetNet with hybrid MA. Most of the research works that can be found nowadays related to NOMA focus on single cell implementations; therefore, we present an analysis that considers a more realistic approach and that can be used as a guide for future 5G network deployments. Since the decision of using either OMA or NOMA for each user in a HetNet depends on how the pairing is done, we also focus on how the pairing selection affects such dimensioning. For this, we compare four generic pairing algorithms, showing their complexity in terms of the runtime, and we propose the use of a cost matrix to help in the pairing selection. The reason for selecting generic pairing methods aims at speeding up the NOMA rollout in 5G networks; in this way, the network operators can rely on such methods while, perhaps, some other optimized methods for each network condition are defined. Through our results we show the impact of NOMA and the pairing method selected on the number of cells needed, users served, and data plans offered Furthermore, we show that a flexible selection of the pairing method, based on the network load conditions, is preferred over a single pairing method.

B. SMALL CELLS AND HetNets

The deployment of cells is needed for network densification purposes; with a massive deployment of cells, network capacity and coverage can be enhanced. Although deploying more macro cells can seem a straightforward solution, finding places for deploying these cells can become increasingly difficult and cost prohibited. There is where small cells play a key role; their reduced size and low power makes them suitable for deployment in places such as lamp posts, traffic lights, and buildings facades. Moreover, the deployment of small cells has become simpler as features as interference, mobility, and software-defined networking (SDN) have been defined by the Third Generation Partnership Project (3GPP) for small cells [4]. New wireless backhaul solutions have also emerged for small cells, facilitating their rollout.

Small cells can be mainly added in hot spots where the data demand is high, by the edge of the cell to benefit the users susceptible to low QoS, and in areas not covered by the macro cells (both outdoor and indoor). At the same time, small cells allow offloading the macro cells, improving the QoS for all the users in the network. With the deployment of small cells, a layer of short-range access points is overlaid on the existing network, allowing this to reduce the distance between the users (UEs) and the base stations (BSs), which results in lower propagation losses, and higher data rates and energy efficiency [8], [9]. Network densification through the deployment of different types of cells essentially leads to HetNets; Fig. 1 shows a typical architecture of a HetNet.

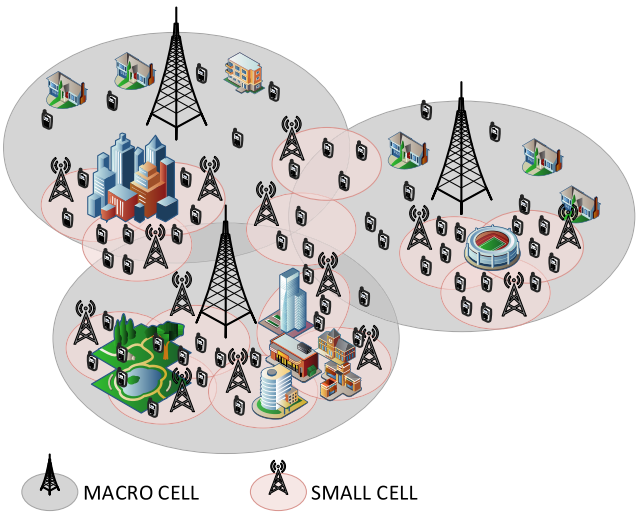


FIGURE 1. HetNet example with hotspots and cells edges covered by small cells.

For in-band deployments of HetNets where all the cells operate at the same frequency, techniques such as inter-cell interference coordination (ICIC), enhanced ICIC (eICIC), coordinated multipoint (CoMP), and enhanced CoMP (eCoMP) have been added by the 3GPP for a more efficient management of inter-cell interference [4]. However, in scenarios where macro cells and small cells operate at different frequencies (out-of-band deployments) the inter-cell interference can be handled with simple interference management methods. In [3], target scenarios for small cells enhancements have been defined for 5G HetNets. The out-of-band implementations in [3] represent one of the biggest advantages/changes for 5G; not only they allow to explore new frequency bands (e.g., millimeter wave bands) to enjoy more and wider spectrum, but also with them the decoupling of the control and user plane (C/U plane split) is possible.

In in-band HetNets, coverage and data services are simultaneously provided by both types of cells, with the control and data plane coupled. This architecture allows for ubiquitous coverage, at the expenses of having all the cells working, even under low load conditions, resulting in a sub-optimal use of resources and energy [10]. On the contrary, in a C/U

plane split architecture, the macro cell is in charge of the control plane, and hence it provides ubiquitous coverage and manages the mobility using the lower frequency bands; it also provides data services for the UEs not covered by small cells. Moreover, the macro cells provide data services to high-speed UEs to avoid frequent handovers in the small cells. The small cells are in charge of the user plane, boosting the capacity by providing high-speed data connections, and more flexible/cost-energy efficient operations in higher frequencies [10]–[12]. Since the propagation losses increase as the frequency increases, high frequencies offer smaller coverage area, thus making them suitable for small cells. With the C/U plane split architecture, the UEs will be simultaneously connected to the macro and the small cells; this dual connectivity allows for a fast handover of the UE to the macro cell in case that the connection to the small cell fails. With this architecture, a new interface will be required through which the macro cell can manage the small cells; this interface will allow the macro cell to activate/deactivate the small cells for energy saving purposes and to participate in the radio resource management to help mitigate the interference [10], [13]. Fig. 2 illustrates a HetNet with C/U plane split architecture.

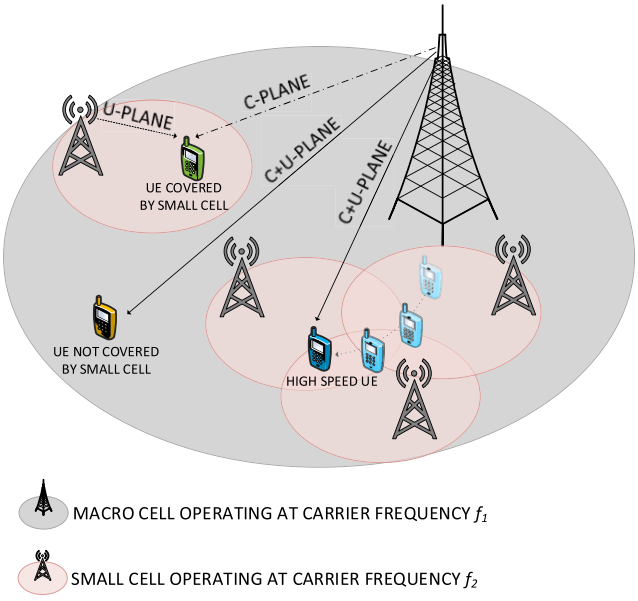


FIGURE 2. HetNet with C/U plane split architecture. High-speed UEs and those not covered by the small cells receive data and control from the macro cell. UEs covered by the small cells receive control from the macro cells and data from the small cells.

Due to the benefits of implementing HetNets, many research works have emerged on the topic. In [14] an overview of HetNets architectures focusing on the capacity and coverage benefits that can be achieved through multilayer and multi-Radio Access Technology (RAT) deployments, is presented. Andrews *et al.* [15] study four different approaches for load balancing in HetNets: relaxed optimization, game theory, Markov decision processes, and range expansion (i.e. biasing); this study shows that although load balancing is still a challenge in HetNets, it offers considerable new flexibility and gain to the system design. The work in [16] focuses on the inter-cell interference in HetNets and evaluates the performance of eICIC techniques. In [17] an extensive research on the technical details and performance gains of HetNets can be found. Moreover, in [18] a framework of a cooperative HetNet for 5G with a particular highlight on the energy efficiency and spectrum efficiency has been studied.

C. NOMA PLUS OMA

Multiple access (MA) techniques are used to allow sharing the available resources among a large number of UEs in the most effective way. As one of the most limited resources in a mobile network is the spectrum, in an MA system different UEs get to simultaneously use the available bandwidth. MA schemes can be broadly classified into two categories: OMA and NOMA [19]. OMA schemes have the advantage of avoiding intra-cell interference but they require careful cell planning to reduce inter-cell interference. The later can be achieved by having sufficient distance between the reused channels, which results in a low spectral efficiency. On the contrary, NOMA schemes are prone to high intra-cell interference, but are robust against fading and inter-cell interference.

In 4G long-term evolution (LTE), orthogonal frequency division multiple access (OFDMA) was chosen for the downlink. The selection of this MA scheme was a key step for increasing the capacity and improving the performance in 4G LTE. Despite the significant enhancements that OFDMA offers, they might not be sufficient to cope with the expected traffic demands for 5G. Therefore, new MA schemes aiming at further increasing the spectral efficiency are highly anticipated. In this regard, NOMA has gained a lot of attention as an MA technique that can boost the capacity of 5G networks, because of its ability to increase the spectral efficiency [5]–[7]. Other benefits of using NOMA include higher cell-edge throughput, relaxed channel feedback, and low transmission latency. Furthermore, with NOMA, a good operating point where both spectrum efficiency and energy efficiency become optimum, can be achieved [20].

Unlike the OMA schemes used in 4G LTE, orthogonality in the resources (e.g., frequency, time, spreading codes) is no longer needed with NOMA. The main idea behind NOMA is to allocate the same frequency channel to two or more multiplexed UEs at the same time. Fig. 3 shows a comparison of users multiplexing between OMA and NOMA for four UEs; here it can be seen that the OMA transmissions are done with full power, while the NOMA ones are done with split power The UEs to be multiplexed in NOMA should be selected in a manner that UEs with high channel conditions can access the resources assigned to UEs with low channel gain, thus achieving a higher spectral efficiency. This is where the advantage of NOMA over OMA schemes used in 4G relies on. At the transmitter side, NOMA uses superposition transmission to join the multiplexed UEs signals; at the

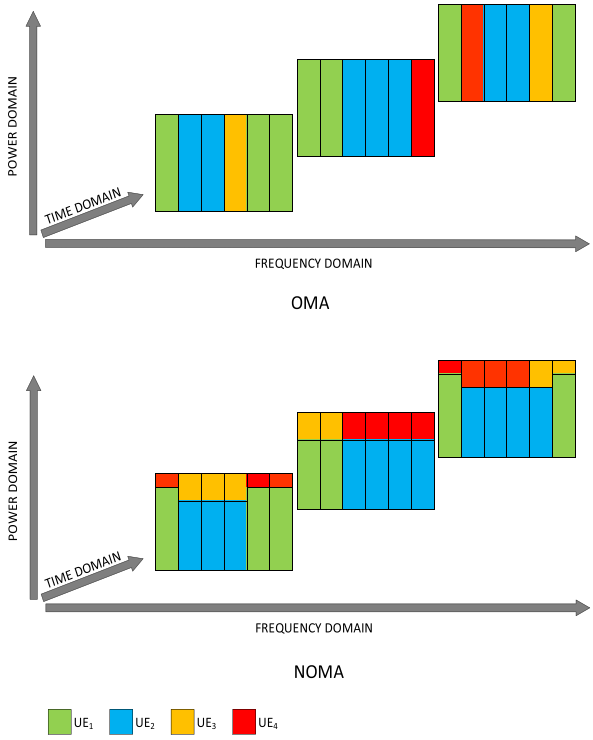


FIGURE 3. Users multiplexing differences between OMA and NOMA.

receiver. successive interference cancellation (SIC) is used to eliminate the multiuser interference [5].

Many research works have already been done regarding NOMA and its performance, challenges, node cooperation, and user pairing. The works in [7] and [20]–[23] present a comprehensive approach to NOMA. Moreover, a NOMA version for the downlink referred to as multiuser superposition transmission (MUST) has been proposed by the 3GPP [24] to be implemented in the future LTE networks. However, the implementation of NOMA in 5G does not mean that it will replace the OMA schemes used nowadays. Depending on the load and the UEs channel conditions, the system might decide to use either OMA or NOMA for each UE. This leads to having a hybrid MA system in 5G, where OMA and NOMA coexist [6], [20], [25]. This coexistence is in accordance with the coexistence of multiple radio access technologies (RATs), which is highly anticipated for 5G. Fig. 4 shows an example of UEs multiplexing in a hybrid MA system.

D. NOMA IN 5G HetNets

The implementation of NOMA in HetNets can offer extended benefits since both technologies share the objective of improving the spectral efficiency. Particularly interesting is the case of NOMA with millimeter wave (mmWave) frequencies. The use of mmWave frequencies (i.e. 30-300 GHz) for the small cells in 5G is a promising implementation [8]. Even though the use of such high frequencies comes with many propagation challenges, its combination with techniques as massive multiple-input-multiple-output (MIMO)

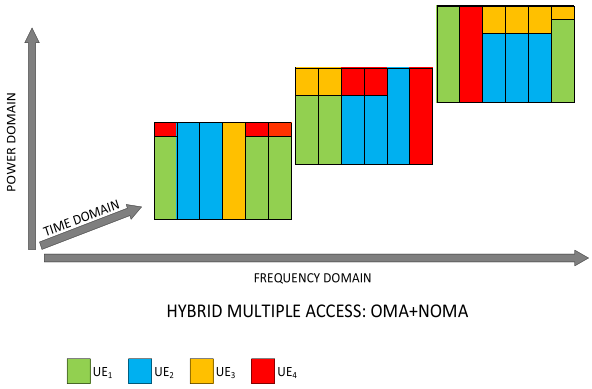


FIGURE 4. Users multiplexing in a hybrid multiple access system, combining OMA and NOMA.

and beamforming strengthens the viability of mmWave in 5G [9], [26], [27]. Nevertheless, the combination of mmWave and NOMA is a challenging aspect. The works in [28]–[32] have researched on the integration of NOMA with mmWave and massive MIMO.

II. THE THEORY BEHIND NOMA

Two main categories of NOMA have been broadly defined in the literature: power-domain NOMA and code-domain NOMA. In the former, the signal of each multiplexed UE is separated in the power domain; the poorer the channel conditions, the higher the power allocated, and vice versa. In the latter, user-specific spreading codes are used to differentiate the multiplexed signals. The work in [20] presents an inside to the most relevant NOMA techniques. In this paper we focus on the power-domain NOMA in the downlink, so from now on we refer to this scheme simply as NOMA.

In NOMA, besides the multiplexing in time and frequency domains, UEs are also multiplexed in the power domain. The principle of NOMA is to select UEs with a high difference in their channel conditions and multiplex them in the same time/frequency resources, but with different levels of transmission power. This allows UEs with high channel conditions to access the resources assigned to UEs with poor channel conditions, hence increasing the spectral efficiency and the system capacity [6]. In the transmitter, signals from the multiplexed UEs are superposed and adaptive power allocation techniques are used to define the power for each UE. The power allocated depends on the channel conditions, the higher the channel gain the higher the power, and vice versa.

Although power-sharing reduces the power allocated to each multiplexed UE, they benefit from being scheduled more often and having access to more bandwidth [33], as shown in Fig. 3. In the receiver side, SIC techniques are used to mitigate the inter-cell interference. The number of UEs that can be multiplexed in the same resources with NOMA is not restricted; however, the inter-cell interference is proportional to the number of UEs. Moreover, the constellation of the superposed signal in the transmitter becomes more complex

as the number of multiplexed UEs increases, posing great challenges on the decoding side and compromising the network performance. Therefore, and for simplicity reasons, for the rest of the paper, we assume only two UEs are multiplexed in the same resources.

A. SUPERPOSITION TRANSMISSION

Superposition transmission is a physical layer technique, first proposed in [34] that allows a single transmitter to simultaneously send a combination of independent signals to several UEs. The transmitted signal after applying superposition techniques for two UEs would be as follows:

$$X = \sqrt{P_1}X_1 + \sqrt{P_2}X_2 \tag{1}$$

with

$$P = P_1 + P_2 = 1 \tag{2}$$

where X_i is the signal corresponding to the UE$_i$'s message, M_i; and P_i is the power ratio for UE$_i$. The difference between the values of P_1 and P_2 should be large enough to guarantee a successful decoding of the superposed signal. The waveform used for the transmissions could be based on orthogonal frequency division multiplexing (OFDM), same as in 4G LTE.

B. SUCCESSIVE INTERFERENCE CANCELLATION

Because of the non-orthogonality of NOMA, interference in the power domain is intentionally added in the transmitter. To mitigate this interference, SIC can be applied [34]. The received signal by UE$_i$ is of the form:

$$Y_i = h_i X + W_i \tag{3}$$

where h_i represents the complex channel coefficients between UE$_i$ and the BS, and $W_i(n)$ represents the Gaussian noise plus inter-cell interference experienced by UE$_i$. The optimal order for decoding the received signal is in the order of the increasing signal strength (i.e. the channel gain normalized by the noise and inter-cell interference) [5]. In this regard, UEs are organized based on their signal strength; so that any UE$_n$ first decodes the strongest signal and removes that from the received combined signal, isolating the desired signal. To better exemplify SIC, let us assume that we have two UEs, UE$_1$ and UE$_2$, and that UE$_2$ is first in the decoding order, hence its signal is the strongest (with more power). In the UE$_2$ receiver, the decoding will go as follows [35]:

1. The message M_2 is decoded from Y_2, treating X_1 as noise. The interference caused by UE$_1$ on UE$_2$ should not significantly affect the performance of UE$_2$, as the power from such interference is likely to be much smaller than the desired signal. This is valid as long as an effective power allocation was performed in the transmitter.

For UE$_1$ the decoding process is more complex and here is where SIC is applied:

1. The message M_2 is decoded from Y_1, treating X_1 as noise. This step is possible because of the fact that the channel gain of UE$_1$ is higher than that of UE$_2$, so as long as the rate of UE$_2$ is within the Shannon limits of its receiver, it will also be within the limits of the UE$_1$ receiver.
2. X_2 is regenerated by using an encoder, and with the knowledge of h_1 and P_2, $h_1\sqrt{P_2}X_2$ is subtracted from Y_1, obtaining:

$$\begin{aligned} Y_1' &= Y_1 - h_1\sqrt{P_2}X_2 \\ &= h_1\sqrt{P_1}X_1 + W_i \end{aligned} \tag{4}$$

3. The message M_1 is decoded from Y_1'.

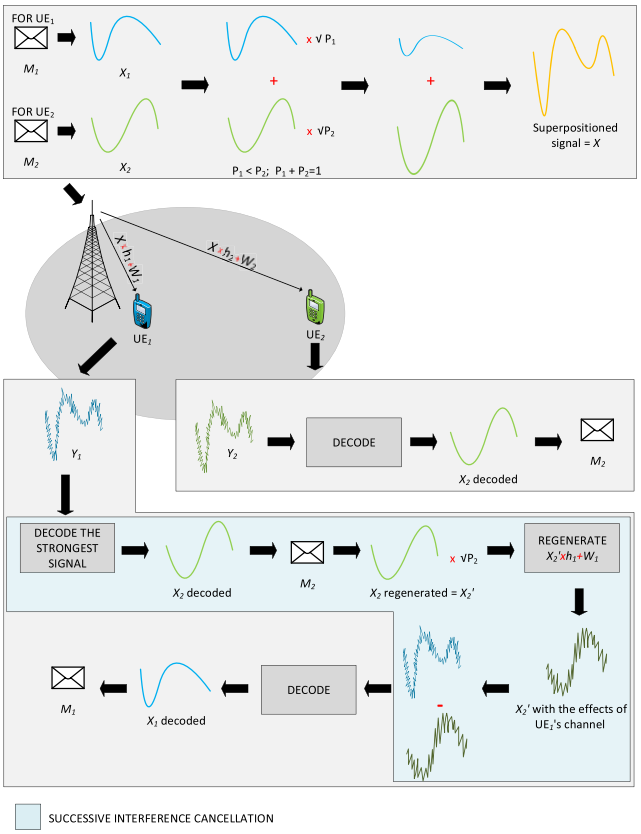


FIGURE 5. Transmission and reception of signals in NOMA for two users.

In [24], several receiver schemes are proposed for NOMA depending on the UE channel conditions. Fig. 5 shows an example of the transmission and reception of NOMA for two UEs. The messages M_1 and M_2 are mapped to the signals X_1 and X_2, respectively. These signals are then scaled according to the values of P_1 and P_2, and summed to generate the supperpositioned signal that is sent to both UEs. During the transmission, each signal is affected by the channel conditions of its respective receiver. Once the signal is received, the far UE, UE$_2$, simply decodes the stronger signal, whereas the near UE, UE$_1$, applies SIC before decoding its signal.

C. DATA RATES

Theoretically, it is known that NOMA offers a bigger capacity region than OMA [20], [21]. Assuming a successful decoding

and no error propagation, the data rates with NOMA for UE_1 and UE_2, can be represented by (5) and (6), respectively:

$$R_1 = \log_2\left(1 + \frac{P_1 |h_1|^2}{N_{o,1}}\right) \tag{5}$$

$$R_2 = \log_2\left(1 + \frac{P_2 |h_2|^2}{P_1 |h_2|^2 + N_{o,2}}\right) \tag{6}$$

where $N_{o,i}$ is the power spectral density of W_i. As the values of R_1 and R_2 depend on the power allocation ratio P_1/P_2, the overall throughput gain of NOMA is tightly related to the power allocation scheme selected. In comparison, for an OMA transmission, the data rates of UE_1 and UE_2 are given by (7) and (8), respectively:

$$R_1 = \alpha \log_2\left(1 + \frac{P_1 |h_1|^2}{\alpha N_{o,1}}\right) \tag{7}$$

$$R_2 = (1-\alpha)\, log_2\left(1 + \frac{P_2 |h_2|^2}{(1-\alpha) N_{o,2}}\right) \tag{8}$$

where α represents the bandwidth assigned to UE_1, with the remaining bandwidth being assigned to UE_2. When using numerical examples, it can be shown that the rate values corresponding to NOMA are considerably higher than those of OMA [5].

D. RESOURCE MANAGEMENT

The appeal of NOMA for 5G networks relies on its more effective utilization of scarce resources (e.g., spectrum) than 4G; therefore, to really exploit the capacity benefits offered by NOMA, resource management must be done in the most effective possible way. In NOMA, there are three resources that must be carefully allocated: power, frequency and time. Since a group of UEs will be assigned to the same frequency channel during the same time, such UEs must be chosen to guarantee that there will be a capacity gain and that resources will not be wasted. Moreover, the power allocation for each multiplexed UE in NOMA must also be carefully chosen to allow the correct decoding of the signals on the receiver side. Both user-pairing and power allocation, are complex processes that require optimization algorithms to allow for the best results with the minimum resources. Some research works have been focused on these two processes, as outlined in the following.

The work in [36] deals with user pairing for two NOMA system: NOMA with fixed power allocation and cognitive-radio-inspired NOMA. Results show that each of these systems exhibits a different behavior when selecting the UEs to be paired, and that the gains of fixed power NOMA over OMA can be further increased by selecting UEs whose channel conditions are more distinctive. In [37], a user pairing and power allocation approach based on a proportional fair (PF) metric is used to achieve a balance between transmission efficiency and user fairness. The proposed scheme offers low computational complexity by deriving the prerequisites for user pairing and avoiding comparison of candidate user pairs. Murti and Shin [38] propose three user pairing methods based on the CQI; results are presented for cases with perfect and imperfect SIC and compared with OMA. Matching theory is proposed in [39] as an approach to optimize user pairing and power allocation in the downlink in a cognitive radio NOMA. Results show that the low complexity proposed algorithm results in a stable matching and outperforms an OMA system. In [40] two user pairing strategies are proposed, where all the users, including those in the middle of the cell who are typically left unpaired, are considered; results show that the proposed algorithm can outperform the near-far pairing, especially in scenarios with imperfect SIC. In [41] a comprehensive review of resource management in NOMA is presented; here the authors propose a resource management framework based on game-theoretic models for power-domain and code-domain NOMA.

III. NOMA IMPLEMENTATION

A. POWER ALLOCATION

The selection of the power ratio in NOMA directly impacts on the UEs data rate and thus in the system performance. Moreover, because of the power-domain multiplexing, the power ratio of one of the multiplexed UE affects the data ratio of not only that UE but also of its pairs, as can be seen in equations (5) and (6).

Despite the research works done on the topic, power allocation in NOMA still remains as an implementation issue. In general, two types of power allocation can be considered for NOMA. One, based on a fixed set of power allocation coefficients, and other based on dynamic power allocation. In this paper, we use the later by implementing the approach suggested in [42], where the power for the NOMA UE with the strongest channel gain is derived from the assumption that the capacity of its paired UE will be the same in NOMA as in NOMA. Equations (9) and (10) are used to estimate the power allocation for UE_1 (UE with higher channel gain) and UE_2, respectively, based on the SINR of UE_2, γ_2:

$$P_1 = \frac{\sqrt{1+\gamma_2} - 1}{\gamma_2}, \text{ with } \gamma_2 > 0 \tag{9}$$

$$P_2 = 1 - P_1 \tag{10}$$

B. USER PAIRING

For the pairing process, UEs are divided into two groups in the scheduler. Group A corresponds to those UEs that have been already selected by the scheduler to transmit in the following subframe; we refer to this as pre-scheduling. Group B corresponds to those UEs that are in need of resources but were not selected to transmit during the pre-scheduling because of lack of resources. The UEs in the groups are not sorted in any particular order. A proportional fair scheduling algorithm is used for the resources assignment and the priority of each UE is assigned according to the following metric $PF_i[t]$ [43]; the UEs with the highest $PF_i[t]$ are scheduled first:

$$PF_i[t] = R_i[t]/S_i[t-1] \tag{11}$$

where t is the subframe number, $R_i[t]$ is the target data rate and it depends on the application in use by UE_i, and $S_i[t-1]$ is its average experienced data rate and can be estimated as:

$$S_i[t] = \frac{t-1}{t} S_i[t-1] + \frac{1}{t} R_i[t] \tag{12}$$

In this paper we evaluate the performance of four generic pairing algorithms for downlink NOMA, aiming at maximizing the system capacity with relatively low complexity. For the first approach, we treat the pairing as an assignment problem, using the Hungarian method [44] to find an optimal solution by which the systems gets the maximum capacity.

For the second approach, we use the Gale-Shapley algorithm [45] to find a stable pairing. Unlike the Hungarian method that finds the optimal solution by minimizing (or maximizing) a cost associated with a set of pairs, the Gale-Shapley algorithm finds an optimal solution based on the stable marriage criterion. Up to date, no work on user pairing for NOMA in HetNets has evaluated the Hungarian method, and the work in [46] proposes an extension of the Gale-Shapley algorithm for user pairing in NOMA. The third algorithm is a simple random pairing, in which UEs from Group A choose the best available pair from Group B. The fourth algorithm is an exhaustive search over all possible pair combinations. For all four algorithms, we first generate a cost matrix that reflects the cost of each possible pair. Table 1 shows an example of such matrix:

TABLE 1. Cost matrix.

		GROUP B				
		UE_{n+1}	UE_{n+2}	UE_{n+3}	UE_j	UE_m
GROUP A	UE_1	$C_{1,n+1}$	$C_{1,n+2}$	$C_{1,n+3}$	$C_{1,j}$	$C_{1,m}$
	UE_2	$C_{2,n+1}$	$C_{2,n+2}$	$C_{2,n+3}$	$C_{2,j}$	$C_{2,m}$
	UE_3	$C_{3,n+1}$	$C_{3,n+2}$	$C_{3,n+3}$	$C_{3,j}$	$C_{3,m}$
	UE_i	$C_{i,n+1}$	$C_{i,n+2}$	$C_{i,n+3}$	$C_{i,j}$	$C_{i,m}$
	UE_n	$C_{n,n+1}$	$C_{n,n+2}$	$C_{n,n+3}$	$C_{n,j}$	$C_{n,m}$

where $C_{i,j}$ is the cost function and is represented by:

$$C_{i,j} = \frac{1}{1 + R_T \Delta_{SINR}}; \quad \text{with } i \le n;\ n \le j \le m \tag{13}$$

with

$$R_T = R_i + R_j; \ \text{with } i \le n; n \le j \le m \tag{14}$$

$$\Delta_{SINR} = \left| SINR_i - SINR_j \right|; \quad \text{with } i \le n;\ n \le j \le m \tag{15}$$

where R_T is the sum of the UE_i and UE_j data rates, according to (5) and (6), and Δ_{SINR} is the difference in the channel gain of UE_i and UE_j. The lower the value of $C_{i,j}$ the higher the pair throughput, R_T. The selection of this cost function aims at facilitating the pairing mainly for the Hungarian and Gale-Shapley algorithms, although it also applies to the other methods. The two former methods are defined to minimize the cost associated with certain pair selection; for this, they give preference to the pairs whose cost is lower (as we explain ahead). Because the rate gain in NOMA is higher as the channel gain difference between the paired UEs increases, as Δ_{SINR} increases so does R_T; thus, approximating the cost function to zero. This means that as the paired UEs offer higher sum rates, a lower cost will be associated with them, giving such pair a higher probability of being chosen during the pairing process. When calculating $C_{i,j}$ the following restrictions apply:

(i) $R_T \ge R_i$, with R_i calculated according to (7). This guarantees that the pairing will result in a capacity gain.

(ii) $I'_{MCS_i} < I'_{MCS_j}$, to guarantee that the UE with high channel gain is the one accessing the resources assigned to the UE with low channel gain.

The values $I'_{MCS_{i,1}}$ and $I'_{MCS_{j,2}}$ correspond to the modulation and coding scheme (MCS) index of UE_i and UE_j, respectively, after the CQI estimation for NOMA.

Because of the inter-cell interference that is intentionally added in NOMA, in a network with hybrid MA the CQI reported by the UE to the BS might not be based on the effective SINR after SIC. Hence, a CQI mismatch between OMA and NOMA transmissions can be expected [32], [42]. A solution for this could be to have all the UEs report the CQI assuming an OMA transmission (i.e. without SIC estimations) and have the BS estimate the CQI for NOMA, in order to select the correct MCS. For this, we use the approximations proposed in [33]:

$$CQI'_i = \frac{P_2 CQI_i}{P_1 CQI_i + 1}; \quad \text{with } i \le n;\ j \le m \tag{16}$$

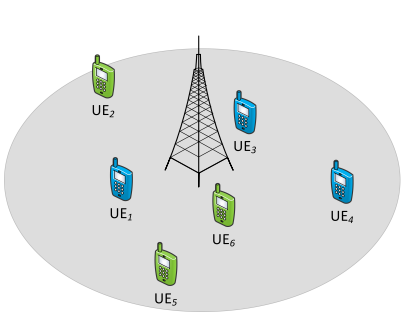
$$CQI'_j = P_1 CQI_i; \quad \text{with } i \le n;\ j \le m \tag{17}$$

where CQI'_i and CQI'_j are the estimated CQIs for NOMA, and CQI_i and CQI_j are the reported CQIs for UE_i and UE_j, respectively. In our model, the CQI reported is estimated based on the SINR and for a block error rate (BLER) of 10%.

The UE pairs that do not fulfill the restrictions (i) and (ii) are considered as non-suitable pairs; thus, in the cost matrix, a cost much higher than the maximum $C_{i,j}$ is assigned, so that such pairs are not considered during the pair selection. Once the cost matrix is generated, we proceed with running the algorithms to find the best pairs that minimize the cost and therefore maximize the system capacity. The number of pairs should be equal to the number of UEs in the group with fewer members, unless there are two or more UEs that can only be paired with the same UE from the opposite group, because of the pairing restrictions set. In case of the latter, fewer pairs than expected will be considered.

For the Hungarian and Gale-Shapley algorithms, if the number of UEs in Group A and Group B are not equal, dummy rows/columns should be used to generate a square matrix, since both algorithms require square matrixes to find the optimal solution.

To exemplify the use of the considered pairing methods in a cell, let us assume that we have a total of six UEs; let us



PRE-SCHEDULED UEs ∈ GROUP A

NOT PRE-SCHEDULED UEs ∈ GROUP B

	SINR (dB)
UE_1	18
UE_2	2
UE_3	16
UE_4	4
UE_5	3
UE_6	20

FIGURE 6. Single cell scenario for user pairing example in NOMA.

also assume that all six UEs need resources in the following subframe and that after the pre-scheduling and the grouping, UEs are divided as shown in Fig. 6.

After calculating the cost of each possible pair according to (13) and assuming that conditions (i) and (ii) are fulfilled, our cost matrix would look as shown in Table 2. The following subsections explain the pairing process and their complexity (i.e., number of iterations) for all four methods considering the example in Fig. 6.

TABLE 2. Cost matrix for user pairing example in NOMA.

		B		
		UE_1	UE_3	UE_4
A	UE_2	0.03	0.04	0.25
	UE_5	0.03	0.03	0.38
	UE_6	0.1	0.05	0.01

C. HUNGARIAN METHOD

The Hungarian method is a combinatorial optimization algorithm used for solving a two-sided one-to-one matching problem. For this method, the problem can be mathematically expressed as:

$$\text{Minimize} \sum_{i=1}^{n} \sum_{j=1}^{m} C_{i,j} x_{ij} \tag{18}$$

where

$$x_{ij} = \begin{cases} 1, & \text{if the UE}_i \text{ is already paired with UE}_j \\ 0, & \text{if the UE}_i \text{ is not paired with UE}_j \end{cases}$$

with the restrictions

(iii) $\sum_{i=1}^{n} x_{ij} = 1; j = 1, 2, \ldots n$, to guarantee that the UE_i only has one pair.

(iv) $\sum_{j=1}^{m} x_{ij} = 1; i = 1, 2, \ldots n$, to guarantee that the UE_j only has one pair.

In Fig. 7 the basic steps of the Hungarian algorithm are applied to the cost matrix in Table 2., until obtaining the final matrix from which pairs are selected by choosing those with $C_{i,j} = 0$. From Fig. 7 we have a total of three pairs

STEP 1
Row reduction: for each row, subtract the minimum value of the row from all the elements in that row. Iterations = 3

		B		
		UE1	UE3	UE4
A	UE2	0	0.01	0.22
	UE5	0	0	0.35
	UE6	0.09	0.04	0

STEP 2
Column reduction: for each column, subtract the minimum value of the column from all the elements in that column. Iterations = 3.

		B		
		UE1	UE3	UE4
A	UE2	0	0.01	0.22
	UE5	0	0	0.35
	UE6	0.09	0.04	0

STEP 3
Zero assignment: starting from the first row, find the rows with only one non-selected zero and select the corresponding pair. Cross out all other zeros in the row and column were the pair was selected. Iterations = 3.

		B		
		UE1	UE3	UE4
A	UE2	0	0.01	0.22
	UE5	0	0	0.35
	UE6	0.09	0.04	0

Selected pair

Crossed-out pair

SOLUTION	COMPLEXITY
UE2 → UE1; UE5 → UE3; UE6 → UE4	TOTAL ITERATIONS = 9

FIGURE 7. Hungarian method application on a user pairing example for NOMA.

as required by the algorithm (i.e., the number of pairs has to be the same as the cost matrix dimension). If after the row and column reduction the number of pairs is not optimal, further steps are taken to optimize the solution. Such steps can be found in [44]. The complexity of this algorithm is calculated as the number of iterations needed to find the optimal pairing. The execution of each step defined by the algorithm is considered an iteration (e.g., each row/column reduction, each zero assignment).

Furthermore, if some of the pairs are those previously defined in the cost matrix as non-suitable, such pairs are omitted during the scheduling and the UEs involved from Group A continue with OMA, whereas the ones from Group B are not scheduled.

D. GALE-SHAPLEY ALGORITHM: PROBLEM FORMULATION

The Gale-Shapley algorithm uses the stable marriage criterion to find stable assignments (pairs). Once the cost matrix is generated, each UE in Group A sorts the UEs in Group B in order of preference, based on the cost functions defined in (11). The lower the cost the higher the preference. The iterative algorithm is then applied to the sorted matrix, during which the UEs from Group A "propose" as a pair to the UEs in Group B. The UEs in the latter either accept (if they are free) or reject (if they are paired and prefer their current pair to the one proposing) the proposition. The solution is said to be stable if, and only if, there exists no UEi and UEj who are not paired with each other but who would both prefer each other over their present partners.

Assuming the same cell scenario as in the previous subsection, we apply the Gale-Shapley algorithm to the matrix in Table 2, until obtaining the final pairing for a total of three

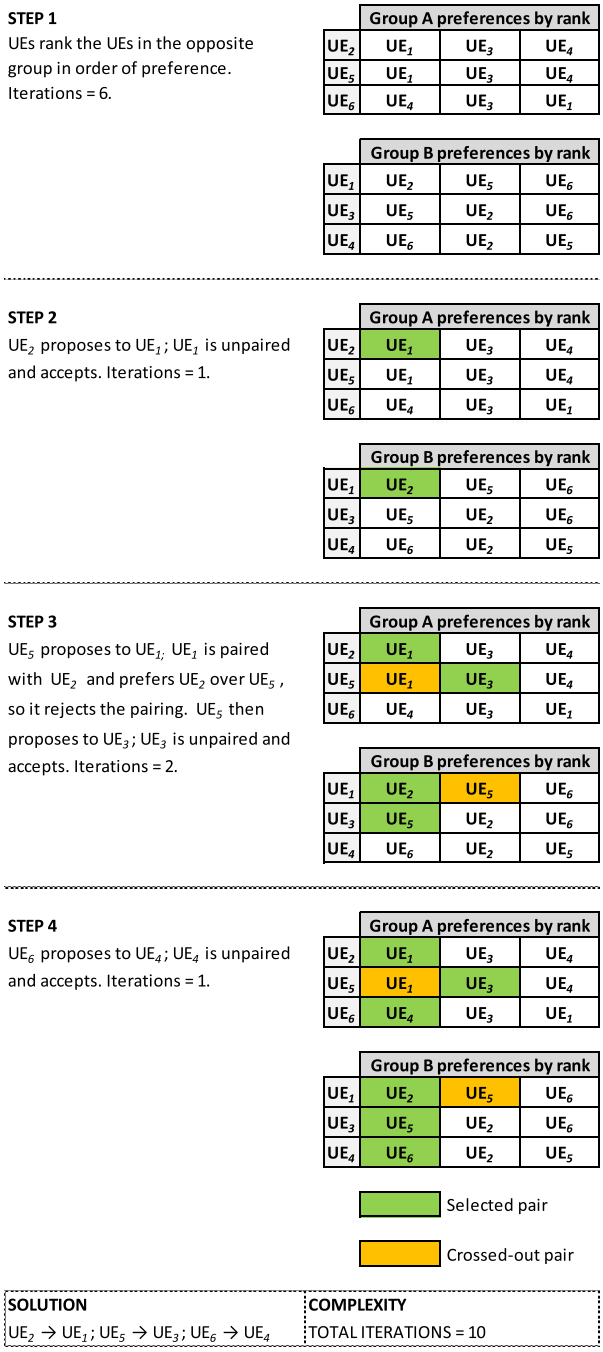


FIGURE 8. Gale-Shapley algorithm application on a user pairing example for NOMA.

pairs. The steps applied are shown in Fig. 8. The complexity of this algorithm is calculated as the iterations for the UEs ranking plus the number of proposals done (accepted and rejected) until the stable solution is found.

Even though the solution provided by the Gale-Shapley algorithm is stable, it is not necessarily the optimal solution. In general, there are several solutions to the pairing when applying this algorithm [47]. The solutions depend on the group that does the proposal. In this regard, and if we follow the dynamic explained above, the stable solution is optimal

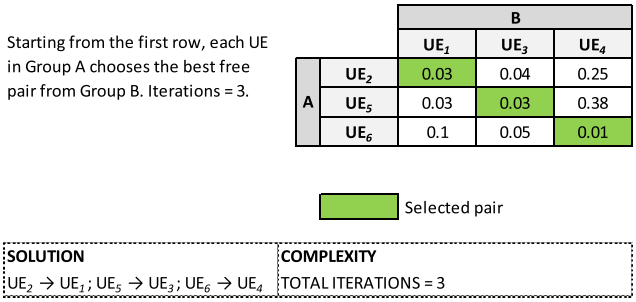


FIGURE 9. Random pairing application on a user pairing example for NOMA.

for the UEs in Group A, but not necessarily for the UEs in Group B. Similarly, if the proposal is done by the UEs in Group B, the solution would be optimal for those UEs. It could be the case, that the stable optimal solution is the same regardless of which group proposes.

Same as with the Hungarian method, if the stable solution considers pairs that have been marked as non-suitable, such pairs are ignored during the scheduling process.

E. RANDOM AND EXHAUSTIVE PAIRING

In the random method, UEs in Group A simply choose the best unpaired UE in Group B, according to the cost of each pair. From the cost matrix in Table 2, the pair choosing process can be seen in Fig. 9. The complexity of this method is calculated as the number of iterations needed until all the pairs have been found.

For the exhaustive pairing, all possible pairs are evaluated to find the combination of pairs that yields the minimum cost. Although the solution from this method is the optimal solution, its complexity makes is computationally expensive. For a cost matrix of size $C\,(n, m)$ with $n \geq m$, a total of $n!/(n-m)!$ iterations are needed to evaluate all the pairs. Each iteration corresponds to the evaluation of one of the possible permutations. When applied to the cost matrix from Table 2, the pairing from the exhaustive search is shown in Fig. 10; six iterations are needed with this method. Although the number of iterations for this example is lower with the exhaustive method, this would not be the case as the dimensions of the cost matrix increase.

From Fig. 7-10 we can see that the same pairing was obtained from all four algorithms. Nevertheless, the process of finding the pairs becomes more complex as more UEs are considered and as some of the pairs do not fulfill the criteria explained above.

F. LOAD BALANCING

The randomness associated with how, when and where the mobile UEs use and demand data from the network, leads to unequal load in neighboring cells. Thus, a cell can be overloaded, while its neighbors have available resources that are not being utilized.

OPTION 1
Cost = 0.07
Iterations = 1

	B		
	UE_1	UE_3	UE_4
UE_2	0.03	0.04	0.25
UE_5	0.03	0.03	0.38
UE_6	0.1	0.05	0.01

OPTION 2
Cost = 0.46
Iterations = 1

	B		
	UE_1	UE_3	UE_4
UE_2	0.03	0.04	0.25
UE_5	0.03	0.03	0.38
UE_6	0.1	0.05	0.01

OPTION 3
Cost = 0.08
Iterations = 1

	B		
	UE_1	UE_3	UE_4
UE_2	0.03	0.04	0.25
UE_5	0.03	0.03	0.38
UE_6	0.1	0.05	0.01

.
.
.

Continue until evaluating all six combinations. Select the one with lowest cost.

SOLUTION	COMPLEXITY
$UE_2 \rightarrow UE_1$; $UE_5 \rightarrow UE_3$; $UE_6 \rightarrow UE_4$	TOTAL ITERATIONS = 6

FIGURE 10. Exhaustive pairing application on a user pairing example for NOMA.

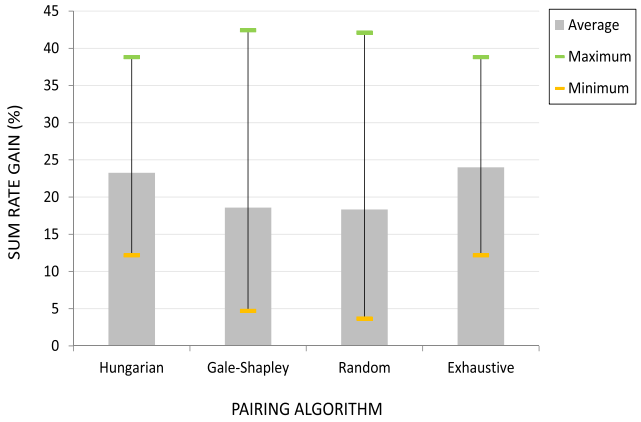


FIGURE 11. Sum rate gain for a single cell for four pairing algorithms for NOMA; OMA is used as the benchmark.

This load imbalance affects the performance of the network, and load balancing techniques can be used to use the resources more efficiently. During these techniques, the BSs communicate to each other and exchange information to compare the cells load.

In this paper, we use the load balancing technique for NOMA proposed in [48]. In this technique, to which we will refer to as LB-NOMA, the balancing is done after the pairs for NOMA have been selected; the purpose is to force handovers of the active UEs that could not be paired (i.e., OMA UEs) and that are located in the overlapping area of two or more cells. This helps to minimize the OMA UEs in the congested cells, thus increasing the system capacity since more resources are available for the paired UEs. The implementation of LB-NOMA comes with many challenges related mainly to the complexity of doing selected load-base handovers and avoiding ping-pong effects for the UEs involved. Nevertheless, we analyze the performance of LB-NOMA in HetNets assuming that these challenges are overcome since they rely on software configurations that can be integrated into the system. Furthermore, with such assumption, we can focus on determining if the gains in the network capacity are significant enough to consider its implementation.

IV. SINGLE CELL PERFORMANCE

For our first performance analysis we consider a highly loaded (i.e., 100% load) single cell with hybrid MA, operating at 2.6 GHz, and with UEs randomly deployed. The pairing algorithms selected for our implementation are compared for this cell. The results are shown for the downlink and use OMA as the benchmark. In Fig. 11 the sum rate gain is shown for all four pairing methods. We can see that the Hungarian method offers a rate gain highly similar to the exhaustive method, with an average of 23.3% and 24%, respectively. The variation for both methods is approximately 12-39%. The results from the Gale-Shapley and random algorithms are also highly similar to each other, with an average gain of 18.3% and 18.5%, respectively; the variation in these methods is wider, which implies a higher uncertainty in the gain that can be obtained. For the Gale-Shapley algorithm, such variation is 4.6-42.4%, whereas for the random algorithm is 3.6-42%.

From Fig. 11 is it clear that the Exhaustive algorithm is the best option for increasing the system capacity, followed by the Hungarian method. Nevertheless, the speed/complexity of the methods should also be considered. Fig. 12 shows the computational complexity in terms of the number of iterations required for each pairing method. We can see from Fig. 12a that the implementation of the exhaustive pairing results in the highest complexity (i.e., number of iterations), with values up to $2E^{+21}$ times higher than the highest complexity for the other three methods. With such complexity, the use of the exhaustive method would likely be time prohibited in a real implementation where every 1 ms a new subframe must be sent. In our implementation, the performance of the exhaustive pairing could be obtained from simulations results for up to 10 pairs, due to software limitations. For higher number of pairs, a combination of simulations results and numerical estimations were used.

In Fig. 12b the complexity of the pairing methods is shown excluding the exhaustive pairing for a better perspective of the performance of the remaining methods. For the Hungarian method, although its complexity is lower than that for the exhaustive pairing, it is on average 85 times higher than for the Gale-Shapley algorithm. The complexity difference between the Hungarian and Gale-Shapley algorithms is due to the fact that with the former an optimal solution must be obtained, whereas with the latter only a stable solution is

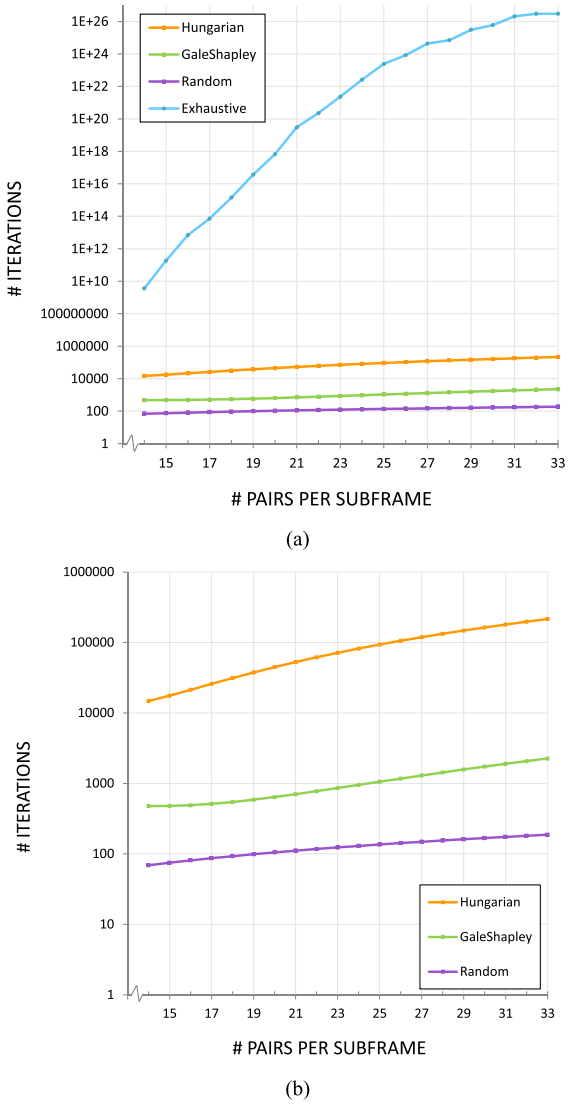


FIGURE 12. Complexity of the pairing algorithms for NOMA, represented as the number of iterations versus the number of pairs per subframe: a) for the four pairing methods considered; b) for the three methods with lower complexity.

needed, which is not necessarily optimal. The selection of one or the other depends on the computational speed and the time constraints. The random algorithm is the one with the lowest complexity, with 9 times fewer iterations needed than with the Gale-Shapley algorithm. Since the capacity gain of the Gale-Shapley over the random algorithm is negligible, the implementation of the random pairing (following the cost function and the pair restrictions proposed above) might offer a better tradeoff between complexity and capacity gain. To further analyze this, a system-level analysis is presented in the next section.

V. NETWORK PERFORMANCE

A. NETWORK MODEL

For evaluating how NOMA affects the downlink capacity dimensioning of a HetNet, we consider a two-tier

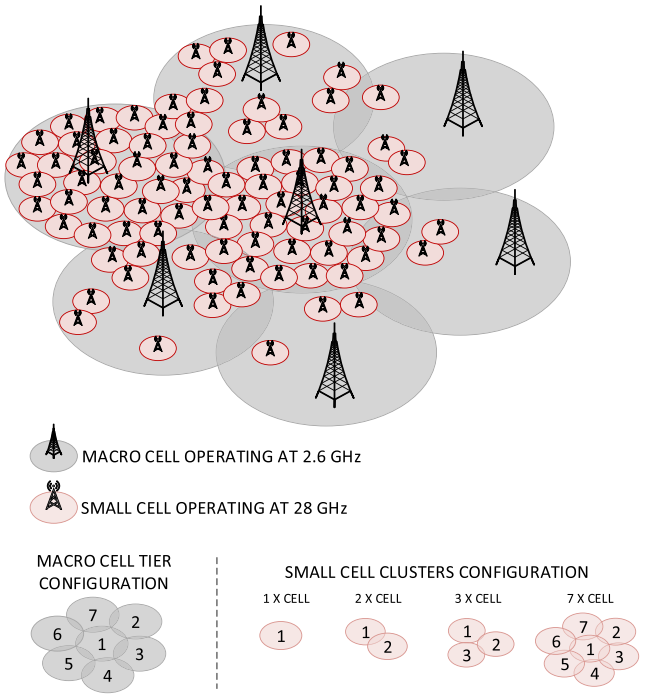


FIGURE 13. Two-tier HetNet model.

out-of-band deployment. The first tier corresponds to the macro cells operating at 2.6 GHz, whereas the second tier corresponds to the small cells operating at 28 GHz; since the study of the coexistence challenges of NOMA and mmWave is outside the scope of this paper, we assume that such design challenges are overcome. This assumption is considered valid since, for mmWave NOMA, solutions related to beamforming techniques such implementation feasible have been proposed in [49] and [50]. The LTE-compatible 5G network model is deployed in MATLAB and consists of a wrap-around cluster model of seven macro cells, with small cells deployed inside their coverage area. The density of small cells on each macro cell depends on the UEs density. Sparse deployment of small cells is used for areas that have identified hotspots, whereas a dense deployment is used for macro cells that are constantly fully loaded. The small cells are modeled as clusters of 7, 3 or 2 cells, or as a single cell. Since the small cells operate at mmWave, the inter-cell interference is rather limited thanks to the high propagation losses at these frequencies [51]. All cells are considered to be located outdoors. Fig. 13 shows the network model, the macro and small cells characteristics are summarized in Table 3 and Table 4, respectively.

For the coverage calculation of the macro cells, an inter-site distance (ISD) of 600 m is used, along with the path loss model 3D-UMa [52]. For the small cells, the ISD is 100 m and the path loss model UMi [53] is used. Table 5 summarizes the parameters used for the link budget calculations and the signal generation.

The UEs in the network are randomly located inside the coverage area of each cell using a normal distribution; the

TABLE 3. Macro cells characteristics.

Macro cell	UE density per Km^2	UEs in cell area	# Small cells deployed inside the macro cell	Type of small cells deployment	% Macro cell area covered by small cells	Average load (%)
1	1200	772	35 (7 x cluster of 7)	Dense: covering most of the macro cell area	0.97	100
2	93	60	3 (1 x cluster2; 1 single)	Sparse: clustered small cells located at the edges	0.08	30
3	62	40	2 (1 x cluster of 2)		0.06	
4	93	60	3 (1 x cluster2; 1 single)		0.08	
5	420	270	10 (2 x cluster of 3; 1 x cluster of 2; 2 single)	Small-area sparse: small clusters and/or isolated cells covering small areas	0.28	60
6	1200	772	35 (7 x cluster of 7)	Dense: covering most of the macro cell area	0.97	100
7	625	402	12 (3 x cluster of 3; 1 x cluster of 2; 1 single)	Small-area sparse: small clusters and/or isolated cells covering small areas	0.33	60

TABLE 4. Small cells characteristics.

Small cell	UE density per Km^2	UEs served by cell	Average load (%)
1	1118	20	100
2	335	6	30
3	335	6	
4	335	6	
5	670	12	60
6	1118	20	100
7	670	12	60

TABLE 5. Parameters for link budget calculation and signal generation.

	MACRO CELL	SMALL CELL
CARRIER FREQUENCY (GHz)	2.6	28
CARRIER BANDWIDTH (MHz)	20	
COMPONENT CARRIERS	5	40
TRANSMISSION POWER (dBm)	43	13
TRANSMITTER ANETNNA GAIN (dBi)	16	10
RECEIVER ANTENNA GAIN(dBi)	0	0
NOISE FIGURE (dB)	7	6
RECEIVER SENSITIVITY (dBm)	-120	-85
PROPAGATION MDOEL	3D-Uma	Umi
ISD (m)	600	100
COVERAGE AREA (km^2)	0.643	0.0179
MULTIPLE ACCESS METHOD	HYBRID: NOMA/LB-NOMA + OMA	
MODULATION SCHEMES	QPSK, 16QAM, 64QAM, 256QAM	
NOMA POWER ALLOCATION	SEE EQUATION 9	

small cells provide the data connections for the UEs inside their coverage area, while the macro cells provide such connections for the rest of the UEs. A total of 2376 UEs are served by the network (Table 3). Four UE profiles are considered, which determine the size and inter-arrival time of the UE's packets: video streaming, FTP file transfer, web browsing and IoT sensors. The work in [54] is used for the characteristics of the first three profiles, whereas the work in [55] is used for the IoT sensors. The profile for each UE is randomly selected while guaranteeing that the average load of the respective serving cell is maintained.

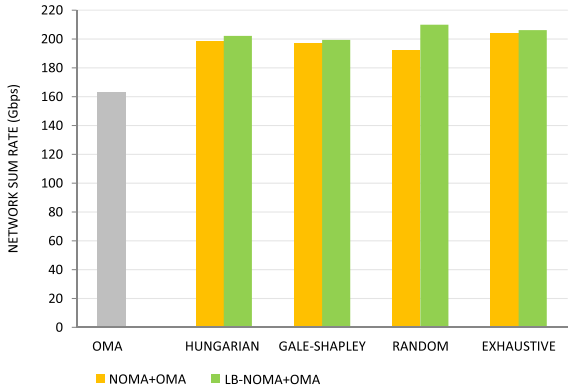


FIGURE 14. Network sum rate comparison for the modeled HetNet.

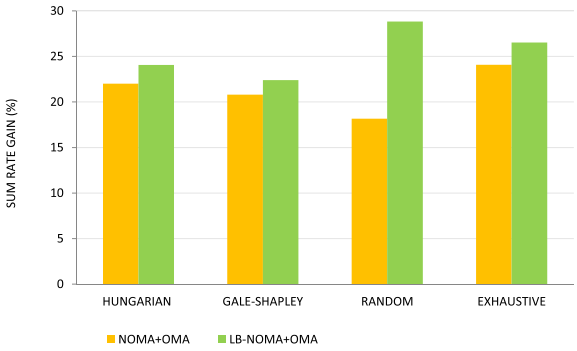


FIGURE 15. Sum rate gain for the modeled HetNet for four pairing algorithms for NOMA; OMA is used as the benchmark.

For the MA scheme, the BS decides between OMA and NOMA depending on the results of the user pairing process. NOMA is applied independently in each network tier.

B. HetNet ANALYSIS AND DIMENSIONING

For this section, the results of our network model in Fig. 13 are analyzed. We consider the four pairing algorithms mentioned above and also compare the performance of the network with LB-NOMA. The results of the system sum rate are presented in Fig. 14, whereas the sum rate gain is shown in Fig. 15 using OMA as the benchmark.

From Fig. 14 and 15 we can verify the better performance that can be achieved by incorporating NOMA in a HetNet. Furthermore, the gain of using LB-NOMA can also be noted, especially for the random pairing. For the NOMA+OMA cases we can see the same trend as in the single cell analysis, which was expected; the Hungarian algorithm offers the highest sum rates and capacity gain after the exhaustive algorithm, achieving an average of 199 Gbps for a gain of 22% and 204 Gbps for a gain of 24%, respectively. In contrast, with the Gale-Shapley algorithm, a sum rate of 196 Gbps is achieved for 21% gain, whereas with a random pairing the sum rate is

approximately 192.5 Gbps for 18% gain. The combination of NOMA with mmWave in the small cells allows having such high rates because of the wide spectrum available and its more effective utilization.

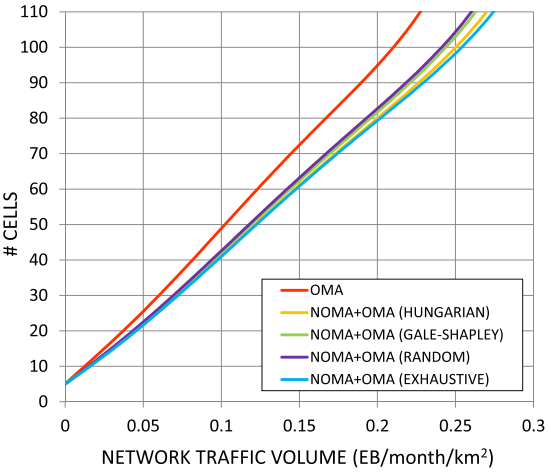
Interestingly, we can see that when LB-NOMA is used, the random method offers the highest capacity gain, with 29% gain corresponding to a sum rate of 210 Gbps; this equals to a gain of 11% because of the use of load balancing in NOMA. To understand this behavior, let us remember that NOMA is more effective as the difference in the channel gain of the paired UEs becomes larger. This typically occurs between UEs located close to the BS and near the edge of the cell. Therefore, when an optimal method is used to find the pairs, the UEs located at the edge of the cell will be chosen with a higher probability, since pairing them with UEs close to the BSs yields the highest system gain. With pairing methods that are not optimal, such as the random and the Gale-Shapley algorithm, the probability of having active UEs at the edge of the cell that are not paired is higher. Hence, when LB-NOMA is implemented, more UEs are likely to be moved to cells that are not fully loaded. This makes the LB-NOMA more efficient when the pairing method fails to choose the best possible pairs. Then, in scenarios where the cells are not equally loaded, the implementation of a simple pairing method such as the random can be chosen along with LB-NOMA. On the contrary, if all cells tend to be fully loaded thus limiting the need for LB-NOMA, optimal pairing methods should be considered.

To now illustrate how the implementation of NOMA affects the capacity dimensioning of a HetNet, let us consider our network model from Fig. 13. For the estimations we use a traffic volume based dimensioning; assuming that during the busy hour the average load of the cells is 50% and that the busy hour carries 15% of the daily traffic, the traffic volume T in GB/month/km2 can be estimated as:

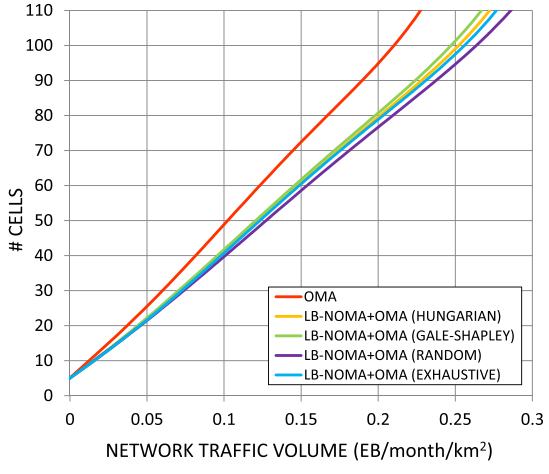
$$T = \frac{\left(\frac{\text{Sum rate [GBps]} \cdot 3600\ [\text{s}] \cdot 50\%}{15\%}\right) \cdot 30\ [\text{days}]}{\text{area [km}^2]} \qquad (19)$$

The results are shown in Fig. 16; the seven macro cells are fixed for every case, and the capacity expansion is done by adding small cells. From these results, we can see the advantages of NOMA from a dimensioning perspective. In a first glance, the most noticeable gain, in terms of the cells needed to support certain traffic volume, is that of including NOMA (independently of the user pairing method) as an MA scheme. This gain is clearer as the traffic volume increases. Because of the massive amount of data expected for 5G networks, the implementation of NOMA is an attractive feature to meet the capacity requirements while minimizing the deployment costs associated.

If our network, for example, needs to handle a volume of 0.2 EB/month/km2, with OMA we would need 98 cells. On the contrary, if we have a hybrid MA system with OMA+NOMA (Fig. 16a) and the pairing is done with the Hungarian algorithm, 81 cells are needed; same



(a)



(b)

FIGURE 16. Number of cells needed versus network traffic volume for OMA and for hybrid MA for four pairing algorithms: a) NOMA+OMA; b) LB-NOMA+OMA.

as with the exhaustive pairing. The highest cell requirement from the hybrid MA cases comes with the use of the Gale-Shapley or random algorithms, with 82 cells needed to support such traffic volume; nevertheless, both methods offer a gain of 16 cells over OMA. Moreover, it is important to highlight that, as shown in Fig. 16, the higher the traffic volume, the higher the gain in the number of cells needed.

For the hybrid MA system with LB-NOMA+OMA (Fig. 16b), the best performance is offered when the random pairing is used, as expected from the results obtained for a single cell scenario shown in Fig. 14-15. For the same traffic volume of 0.2 EB/month/km2, 75 cells are needed for the random pairing. The remaining three methods, each needs 81 cells; this represents little to no improvement compared to their equivalents in the NOMA+OMA cases, especially for scenarios with lower traffic volume. Thus, the benefits of LB-NOMA highly depend on the chosen pairing method. When close-to-optimal methods are used, the space left for improvements with LB-NOMA is limited. On the contrary, simpler

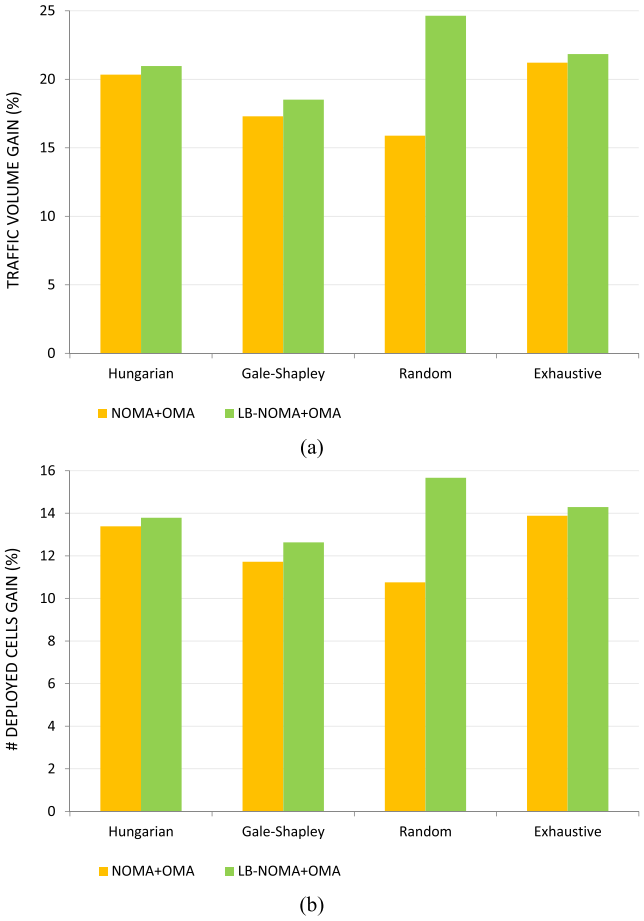


FIGURE 17. Gain for the modeled HetNet for four pairing algorithms for NOMA; OMA is used as the benchmark: a) Traffic volume gain; b) Number of deployed cells gain.

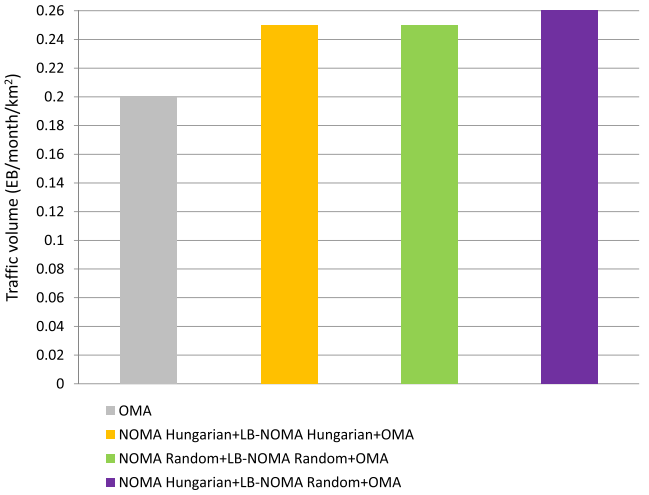


FIGURE 18. Traffic volume that can be supported by the modeled HetNet, with 100 cells deployed for OMA and for NOMA.

algorithms can be used if their weaknesses are balanced with other optimization techniques, such as load balancing.

In Fig. 17 the traffic volume gain and the gain in the number of deployed cells are shown, using OMA as the benchmark. For a NOMA+OMA implementation, the use of the Hungarian method is preferred, since it offers an average gain of 20% in the traffic volume (Fig. 17a) that can be supported and its complexity is lower than the exhaustive method, which offers a 21% traffic volume gain. The Gale-Shapley algorithm offers a 17% gain, whereas the gain for the random method is 16%. On the contrary, for a LB-NOMA+OMA implementation, the random method offers the highest gain in the traffic volume supported, with 24.6%.

For the number of cells needed, we can see from Fig. 17b that for NOMA+OMA the highest gain in the number of cells deployed (a gain meaning fewer cells needed) is achieved with the exhaustive method, with a gain of 14%. However, considering the complexity of the exhaustive method, the Hungarian method could offer a better trade-off between complexity and gain, with 13% saving in cells needed. For LB-NOMA+OMA the highest gain is achieved with the random method, with 15.6%. The gain in the number of deployed cells can be translated into a gain in the deployment cost of the network; in this respect, if 15.6% fewer cells are needed to cope with the capacity demand, roughly 15.6% can be saved from the deployment cost, while maintaining the network revenues.

The selection of the pairing method in a hybrid MA HetNet could then be flexible and subject to the load conditions of the cell and its neighbors. The BSs could choose the best method for pairing in NOMA according to the network conditions. This will allow using simpler and faster, but less efficient, pairing algorithms when the load in the cells is unequal, and compensate the inefficiency of the pairing by using LB-NOMA. The optimal or more efficient algorithms could then be reserved for cases where LB-NOMA is not applied, either because all cells have a similar load or because such feature is not available. By having this flexibility in the implementation of NOMA, the network capacity can be improved while lowering the deployment costs.

For example, in our network, the Hungarian method could be selected for NOMA, whereas the random method could be used for cases when LB-NOMA is beneficial (unequally loaded cells). With such implementation, and assuming that 100 cells can be deployed (7 macro cells plus 93 small cells), an average network traffic volume of 0.26 EB/month/km^2 can be supported, as shown in Fig. 18. In contrast, if only the Hungarian or the random method is used, the same number of cells can handle 0.25 EB/month/km^2. This difference of 4% is only due to the flexible choice of the pairing method and directly translates into a 4% gain in the network revenues, since either more UEs or higher data plans can be supported.

Moreover, either combination of hybrid MA offers significantly higher capacity than the use of only OMA, with which 0.20 EB/month/km^2 can be supported with the same 100 cells; this corresponds to capacity gains between 25-30% because of the use of NOMA. For a simple revenue estimation, we can refer to Table 6, where we consider

TABLE 6. Simple network revenue estimation based only on the end users revenues, for three combinations of multiple access.

Multiple Access	OMA	NOMA + OMA	NOMA/LB-NOMA + OMA		
Pairing Method	N/A	Hungarian	NOMA: Hungarian LB-NOMA: Random	Only Hungarian	Only Random
# Cells	100				
Traffic volume (EB/month/km^2)	0.2	0.25	0.26	0.25	0.25
Network area (km^2)*	4.501				
Price per GB ($)	2				
Revenue (MM$/month)	1.80	2.25	2.34	2.25	2.25
Revenue gain (Benchmark: OMA)	N/A	25	30	25	25

TABLE 7. Simple network user capacity estimation, for three combinations of multiple access.

Multiple Access	OMA	NOMA + OMA	NOMA/LB-NOMA + OMA		
Pairing Method	N/A	Hungarian	NOMA: Hungarian LB-NOMA: Random	Only Hungarian	Only Random
# Cells	100				
Traffic volume (EB/month/km^2)	0.2	0.25	0.26	0.25	0.25
Network area (km^2)	4.501				
GB/month/UE (for 2299 simulated UEs)	379	474	493	474	474
UEs/month (for 30 GB/UE/month)	45,010	56,263	58,513	56,263	56,263

100 cells and the pairing methods with the best tradeoff between complexity/capacity gain for our network (i.e., Hungarian for NOMA and random for LB-NOMA). Only end users revenues are considered and we assume that the price of each GB/month is $2 (example value estimated from data plans commercially offered nowadays). The highest revenues correspond to the hybrid multiple access with flexible pairing method, with 2.34 million dollars per month, offering a 4% gain over the other hybrid MA access considered and a 30% gain over OMA. Furthermore, if we consider the 2376 UEs simulated, an average of 493 GB/month/UE can be offered with the hybrid MA and flexible pairing as shown in Table 7, being the highest monthly data allowance. Since such a huge amount of data will be likely too high for the average monthly consumption, we can see that for a data plan of 30 GB/UE/month, up to 58,513 UEs could be served by the network, versus 56,263 and 45,010 for the other hybrid MA options and OMA, respectively. These numbers result interesting considering the high data demands and increase number of connected devices connected that are expected for 5G networks.

VI. CONCLUSIONS

In this paper, we have analyzed, from a system-level perspective, the performance of a HetNet with hybrid MA. Our network model consists of an out-of-band deployment of macro and small cells, where NOMA and OMA coexist. For NOMA we made particular emphasis on the pairing process, comparing the performance of four generic pairing methods: Hungarian method, Gale-Shapley algorithm, random pairing, and exhaustive pairing. Furthermore, we considered the use of load balancing techniques with NOMA (LB-NOMA) to further increase the overall network capacity. Our results show the clear capacity benefits of hybrid MA over OMA in HetNets. More interestingly, our results also showed the impact of the efficiency of the pairing method and the use of load balancing techniques for NOMA in the overall network capacity. The use of optimal or close-to-optimal pairing methods offers the higher capacity gains (22-24%) in cases where load balancing techniques are not used. On the contrary, if LB-NOMA is used, simpler pairing methods can offer a higher gain (approximately 29%); that is because the inefficiency in choosing the best pairs can be compensated through load balancing techniques. We showed in our dimensioning results that hybrid MA with LB-NOMA+OMA and with random pairing requires significantly fewer cells deployed to offer the same capacity than the other combinations considered. For a network traffic volume of 0.2 EB/month/km2, with OMA 98 cells are needed, with NOMA+OMA and exhaustive pairing 81 cells are needed, whereas with LB-NOMA+OMA and random pairing 75 cells are needed. These results lead us to consider a flexible selection of the pairing method for NOMA depending on the load conditions of the network. For scenarios where all cells are equally loaded and thus load balancing techniques are not effective, the selection of a close-to-optimal (typically more complex) pairing method is preferred. On the contrary, in scenarios where cells are unequally loaded, the selection of simpler and faster pairing methods combined with load balancing techniques could offer a better performance for the network, in terms of capacity and pairing complexity in NOMA. In our network, for 100 cells deployed the combinations of the Hungarian and the random method, for NOMA and LB-NOMA respectively, allows supporting 4% higher network traffic volume, than if either of the two methods is exclusively used regardless of the network load conditions. This 4% gain reflects directly as a revenue gain for the network since either more UEs can be supported or higher data rate plans can be offered.

The analysis of HetNets with hybrid MA combining OMA and NOMA is still on early stages, hence research work on how the resource management on such networks influences the dimensioning and deployment of 5G networks is highly anticipated. Further research on how to select the appropriate pairing method depending on the network conditions should be done, considering variables such as the changes in the cells load, e.g., during the peak hours, or during night hours in commercial and business areas.

REFERENCES

[1] *Ericsson Mobility Report*, Ericsson, Stockholm, Sweden, 2017.
[2] J. Zhao, Y. Liu, K. K. Chai, A. Nallanathan, Y. Chen, and Z. Han, "Spectrum allocation and power control for non-orthogonal multiple access in HetNets," *IEEE Trans. Wireless Commun.*, vol. 16, no. 9, pp. 5825–5837, Sep. 2017.
[3] *Scenarios and Requirements for Small Cell Enhancements for E-UTRA and E-UTRAN (Release 14)*, document TR 36.93, V14.0.0, 3GPP, 2017.

[4] H. Holma, A. Toskala, and J. Reunanen, *LTE Small Cell Optimization: 3GPP Evolution to Release 13*. Hoboken, NJ, USA: Wiley, 2016.
[5] 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 Veh. Technol. Conf. (VTC Spring)*, Jun. 2013, pp. 1–5.
[6] Z. Ding *et al.*, "Application of non-orthogonal multiple access in LTE and 5G networks," *IEEE Commun. Mag.*, vol. 55, no. 2, pp. 185–191, Feb. 2017.
[7] Y. Yuan *et al.*, "Non-orthogonal transmission technology in LTE evolution," *IEEE Commun. Mag.*, vol. 54, no. 7, pp. 68–74, Jul. 2016.
[8] T. S. Rappaport *et al.*, "Millimeter wave mobile communications for 5G cellular: It will work!" *IEEE Access*, vol. 1, pp. 335–349, 2013.
[9] T. E. Bogale and L. B. Le, "Massive MIMO and mmWave for 5G wireless HetNet: Potential benefits and challenges," *IEEE Veh. Technol. Mag.*, vol. 11, no. 1, pp. 64–75, Mar. 2016.
[10] H. A. U. Mustafa, M. A. Imran, M. Z. Shakir, A. Imran, and R. Tafazolli, "Separation framework: An enabler for cooperative and D2D communication for future 5G networks," *IEEE Commun. Surveys Tuts.*, vol. 18, no. 1, pp. 419–445, 1st Quart., 2016.
[11] NTT DOCOMO. (2014). *DOCOMO 5G White Paper—5G Radio Access: Requirements, Concept and Technologies*. [Online]. Available: https://www.nttdocomo.co.jp/english/corporate/technology/whitepaper_5g/
[12] A. Mohamed, O. Onireti, M. A. Imran, A. Imran, and R. Tafazolli, "Control-data separation architecture for cellular radio access networks: A survey and outlook," *IEEE Commun. Surveys Tuts.*, vol. 18, no. 1, pp. 446–465, 1st Quart., 2015.
[13] C. Dehos, J. L. González, A. De Domenico, D. Kténas, and L. Dussopt, "Millimeter-wave access and backhauling: The solution to the exponential data traffic increase in 5G mobile communications systems?" *IEEE Commun. Mag.*, vol. 52, no. 9, pp. 88–95, Sep. 2014.
[14] S.-P. Yeh, S. Talwar, G. Wu, N. Himayat, and K. Johnsson, "Capacity and coverage enhancement in heterogeneous networks," *IEEE Wireless Commun.*, vol. 18, no. 3, pp. 32–38, Jun. 2011.
[15] J. Andrews, S. Singh, Q. Ye, X. Lin, and H. Dhillon, "An overview of load balancing in HetNets: Old myths and open problems," *IEEE Wireless Commun.*, vol. 21, no. 2, pp. 18–25, Apr. 2014.
[16] D. Lopez-Perez, I. Guvenc, G. de la Roche, M. Kountouris, T. Q. S. Quek, and J. Zhang, "Enhanced intercell interference coordination challenges in heterogeneous networks," *IEEE Wireless. Commun.*, vol. 18, no. 3, pp. 22–30, Jun. 2011.
[17] R. Q. Hu and Y. Qian, *Heterogeneous Cellular Networks*. Hoboken, NJ, USA: Wiley, 2013.
[18] R. Q. Hu and Y. Qian, "An energy efficient and spectrum efficient wireless heterogeneous network framework for 5G systems," *IEEE Commun. Mag.*, vol. 52, no. 5, pp. 94–101, May 2014.
[19] P. Wang, J. Xiao, and L. Ping, "Comparison of orthogonal and non-orthogonal approaches to future wireless cellular systems," *IEEE Veh. Technol. Mag.*, vol. 1, no. 3, pp. 4–11, Sep. 2006.
[20] S. M. 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 Commun. Surveys Tuts.*, vol. 19, no. 2, pp. 721–742, 2nd Quart., 2017.
[21] A. Benjebbour, K. Saito, A. Li, Y. Kishiyama, and T. Nakamura, "Non-orthogonal multiple access (NOMA): Concept, performance evaluation and experimental trials," in *Proc. Int. Conf. Wireless Netw. Mobile Commun. (WINCOM)*, Oct. 2015, pp. 1–6.
[22] Z. Ding, X. Lei, G. K. Karagiannidis, R. Schober, J. Yuan, and V. Bhargava, "A survey on non-orthogonal multiple access for 5G networks: Research challenges and future trends," *IEEE J. Sel. Areas Commun.*, vol. 35, no. 10, pp. 2181–2195, Oct. 2017.
[23] L. Dai, B. Wang, Y. Yuan, S. Han, C.-L. I, 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.
[24] *Study on Downlink Multiuser SuperposItion Transmission (MUST) for LTE (Release 13)*, 3GPP, document TR 36.859, V13.0.0. 2016.
[25] A. S. Marcano and H. L. Christiansen, "A novel method for improving the capacity in 5G mobile networks combining NOMA and OMA," in *Proc. IEEE 85th Veh. Technol. Conf. (VTC Spring)*, Jun. 2017, pp. 1–5.
[26] F. Rusek *et al.*, "Scaling up MIMO: Opportunities and challenges with very large arrays," *IEEE Signal Process. Mag.*, vol. 30, no. 1, pp. 40–60, Jan. 2013.
[27] A. L. Swindlehurst, E. Ayanoglu, P. Heydari, and F. Capolino, "Millimeter-wave massive MIMO: The next wireless revolution?" *IEEE Commun. Mag.*, vol. 52, no. 9, pp. 56–62, Sep. 2014.
[28] Z. Ding and H. V. Poor, "Design of massive-MIMO-NOMA with limited feedback," *IEEE Signal Process. Lett.*, vol. 23, no. 5, pp. 629–633, May 2016.
[29] D. Zhang, Z. Zhou, C. Xu, Y. Zhang, J. Rodriguez, and T. Sato, "Capacity analysis of NOMA with mmWave massive MIMO systems," *IEEE J. Sel. Areas Commun.*, vol. 35, no. 7, pp. 1606–1618, Jul. 2017.
[30] 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.
[31] Y. Liu, Z. Qin, M. Elkashlan, Y. Gao, and A. Nallanathan, "Non-orthogonal multiple access in massive MIMO aided heterogeneous networks," in *Proc. IEEE Global Commun. Conf. (GLOBECOM)*, Dec. 2016, pp. 1–6.
[32] A. S. Marcano and H. L. Christiansen, "Performance of non-orthogonal multiple access (NOMA) in mmwave wireless communications for 5G networks," in *Proc. IEEE ICNC*, Jan. 2017, pp. 969–974.
[33] A. Benjebbour, A. Li, Y. Kishiyama, H. Jiang, and T. Nakamura, "System-level performance of downlink NOMA combined with SU-MIMO for future LTE enhancements," in *Proc. Globecom Workshops (GC Wkshps)*, Dec. 2014, pp. 706–710.
[34] T. M. Cover, "Broadcast channels," *IEEE Trans. Inf. Theory*, vol. IT-18, no. 1, pp. 2–14, Jan. 1972.
[35] S. Vanka, S. Srinivasa, Z. Gong, P. Vizi, K. Stamatiou, and M. Haenggi, "Superposition coding strategies: Design and experimental evaluation," *IEEE Trans. Wireless Commun.*, vol. 11, no. 7, pp. 2628–2639, Jul. 2012.
[36] Z. Ding *et al.*, "Impact of user pairing on 5G nonorthogonal multiple-access downlink transmissions," *IEEE Trans. Veh. Technol.*, vol. 65, no. 8, pp. 6010–6023, Aug. 2016.
[37] F. Liu, P. Mähönen, and M. Petrova, "Proportional fairness-based user pairing and power allocation for non-orthogonal multiple access," in *Proc. IEEE 26th Annu. Int. Symp. Pers., Indoor Mobile Radio Commun. (PIMRC)*, Aug./Sep. 2015, pp. 1127–1131.
[38] F. W. Murti and S. Y. Shin, "User pairing schemes based on channel quality indicator for uplink non-orthogonal multiple access," in *Proc. 9th Int. Conf. Ubiquitous Future Netw. (ICUFN)*, Jul. 2017, pp. 225–230.
[39] W. Liang, Z. Ding, Y. Li, and L. Song, "User pairing for downlink non-orthogonal multiple access networks using matching algorithm," *IEEE Trans. Commun.*, vol. 65, no. 12, pp. 5319–5332, Dec. 2017.
[40] M. B. Shahab, M. Irfan, M. F. Kader, and S. Y. Shin, "User pairing schemes for capacity maximization in non-orthogonal multiple access systems," *Wireless Commun. Mobile Comput.*, vol. 16, no. 17, pp. 2884–2894, 2016.
[41] L. Song, Y. Li, Z. Ding, and H. V. Poor, "Resource management in non-orthogonal multiple access networks for 5G and beyond," *IEEE Netw.*, vol. 31, no. 4, pp. 8–14, Jul./Aug. 2017.
[42] Y. Zhu, H.-J. E. Kwon, H. Jung, U. Kumar, and J.-K. J. K. Fwu, "Non-orthogonal multiple access (NOMA) wireless systems and methods," U.S. Patent EP3 138 227 A1, Mar. 8, 2017.
[43] G. Miao, J. Zander, K. W. Sung, and S. B. Slimane, *Fundamentals of Mobile Data Networks*. Cambridge, U.K.: Cambridge Univ. Press, 2016.
[44] H. W. Kuhn, "The Hungarian method for the assignment problem," *Naval Res. Logistics*, vol. 2, nos. 1–2, pp. 83–97, 1955.
[45] D. Gale and L. S. Shapley, "College admissions and the stability of marriage," *Amer. Math. Monthly*, vol. 69, no. 1, pp. 1–15, 1962.
[46] S. Zhang, B. Di, L. Song, and Y. Li, "Radio resource allocation for non-orthogonal multiple access (NOMA) relay network using matching game," in *Proc. IEEE Int. Conf. Commun. (ICC)*, May 2016, pp. 1–6.
[47] D. G. McVitie and L. B. Wilson, "The stable marriage problem," *Commun. ACM*, vol. 14, no. 7, pp. 486–490, 1971.
[48] A. S. Marcano and H. L. Christiansen, "System-level performance of 1162 C-NOMA: A cooperative scheme for capacity enhancements in 5G mobile networks," in *Proc. IEEE 86th Veh. Technol. Conf. (VTC Fall)*, Sep. 2017, pp. 1–5.
[49] Z. Ding, P. Fan, and H. V. Poor, "Random beamforming in millimeter-wave NOMA networks," *IEEE Access*, vol. 5, pp. 7667–7681, 2017.
[50] Z. Xiao, L. Dai, Z. Ding, J. Choi, and P. Xia. (2017). "Millimeter-wave communication with non-orthogonal multiple access for 5G." [Online]. Available: https://arxiv.org/abs/1709.07980
[51] T. S. Rappaport, R. W. Heath, Jr., C. R. Daniels, and N. J. Murdock, *Millimeter Wave Wireless Communications*. Upper Saddle River, NJ, USA: Prentice-Hall, 2014.

[52] *Study on 3D Channel Model for LTE (Release 12)*, document TR 36.873, V12.3.0, 3GPP, 2016.
[53] *Study on channel model for frequency spectrum Above 6 GHz (Release 14)*, document TR 38.900, V14.10, 3GPP, 2016.
[54] R. Porat *et al.*, *11ax Evaluation Methodology*, document IEEE 802.11-14/0571r12, Jan. 2016. [Online]. Available: https://mentor.ieee.org/802.11/dcn/14/11-14-0571-12-00ax-evaluation-methodology.docx
[55] C. Mavromoustakis, G. Mastorakis, and J. M. Batalla, *Internet of Things (IoT) in 5G Mobile Technologies*, 1st ed. Cham, Switzerland: Springer, 2016.

ANDREA S. MARCANO received the B.Sc. degree in telecommunications engineering from the Andrés Bello Catholic University, Caracas, Venezuela, in 2011, and the M.Sc. degree in electronic engineering from Simón Bolívar University, Caracas, in 2013, with a focus on mobile communications. She is currently pursuing the Ph.D. degree with the Technical University of Denmark. She was involved in the dimensioning and optimization of mobile networks and her current research is focused on 5G mobile networks.

HENRIK L. CHRISTIANSEN received the M.Sc.E.E. and Ph.D. degrees specializing in telecommunications from the Technical University of Denmark. He is currently an Associate Professor in mobile communication with the Technical University of Denmark. He has several years of experience from the telecom industry. His main areas of research are mobile network architectures, mobile fronthaul, and backhaul networks.