sensors

MDPI

Article

Non-Orthogonal Multiple Access for Unicast and Multicast D2D: Channel Assignment, Power Allocation and Energy Efficiency

Mariem Hmila *, Manuel Fernández-Veiga, Miguel Rodríguez-Pérez and Sergio Herrería-Alonso

atlanTTic Laboratory, Faculty of Telecommunications Engineering, Universidade de Vigo, 36310 Vigo, Spain; mveiga@det.uvigo.es (M.F.-V.); miguel@det.uvigo.gal (M.R.-P.); sha@det.uvigo.es (S.H.-A.)

* Correspondence: meriame@det.uvigo.es

Abstract: Non-orthogonal multiple access (NOMA) techniques have emerged in the past years as a solution to approximate the throughput performance of wireless communications systems to their theoretical capacity region. We consider in this paper an optimization-based model for multicast device-to-device (MD2D) communications where the channels are not orthogonal and may be (partially or fully) shared among the transmitters in each cluster. This setting leads naturally to the introduction of NOMA transmitters and receivers who use successive interference cancellation (SIC) to separate the superposed signals. To analyze the role of NOMA in MD2D, its performance impact, potential performance gains and possible shortcomings, we formulate a model that includes SIC operations in the decoders, so that higher rates can be attained when several sources transmit on the same channel(s). We also investigate the energy efficiency of the network (global and max-min) through a dynamic power control algorithm and present a centralized and a semi-distributed solution to these optimization problems. Through numerical simulations, we show that NOMA is able to improve both the sum-rate and the max-min rate of a MD2D network even from a small degree of resource sharing. Furthermore, these gains also improve the global energy efficiency on the network, but not always the max-min energy efficiency of the devices.

Keywords: multicast device-to-device communication; 5G and beyond; non-orthogonal communications

check for updates

Citation: Hmila, M.; Fernández-Veiga, M.; Rodríguez-Pérez, M.; Herrería-Alonso, S. Non-Orthogonal Multiple Access for Unicast and Multicast D2D: Channel Assignment, Power Allocation and Energy Efficiency. *Sensors* **2021**, *21*, 3436. https://doi.org/10.3390/s21103436

Academic Editor: Redha Radaydeh

Received: 22 April 2021
Accepted: 11 May 2021
Published: 14 May 2021

1. Introduction

5G and beyond networks are poised to achieve spectrum efficiency increases from five to fifteen times compared to current 4G technology and densities around a million devices per square kilometer [1]. Multicast Device-to-Device (MD2D) communication is a conceptually simple technique that allows users in close proximity to communicate directly without the intervention of a third party, such as a base station (BS) or an access point. These short-range communications incur lower latencies and require less transmission power or, conversely, achieve higher transmission rates for the same power, thus increasing energy efficiency. As a result, both the energy efficiency and the system capacity are improved, and MD2D can contribute significantly to meet the requirements of vast 5G infrastructures.

However, in underlay MD2D communications, simultaneous transmissions over the same resource blocks (RBs) increase interference and may limit the network performance in both metrics, sum-rate and energy efficiency. An engineered allocation of resources (i.e., frequency channels and power levels) is thus necessary to mitigate interference and maximize network performance, ideally in a distributed manner, so as to also minimize the overhead and the computation load at any central site. Since the spectrum bands in MD2D are heavily reused by multiple transmitters, including the normal cellular users (CUs), the channel between these and the receivers in different clusters/groups can be modeled as a multi-user broadcast channel (MU-BC) using orthogonal multiple access

(OMA) by the transmitters. Interference management in this setting has been extensively investigated both through separate or joint channel allocation and power control techniques, i.e., treating the interference at the receivers as noise (e.g., [2–5]). In contrast, with Non-Orthogonal Multiple Access (NOMA) [6,7], receivers are able to separate superposed signals distinguishing them in the power domain, by forcing some users to use Successive Interference Cancellation (SIC). By SIC, a user fully and sequentially decodes the signals intended for other weaker users, subtracts these from the received signal, and finally recovers its own signal with much lower noise. Thus, NOMA offers high prospects to suppress interference at the receivers, and has been thoroughly studied in the 5G design stage (and since much earlier in the information theory literature [8]) with remarkable success for achieving high efficiency [9–14]. Therefore, NOMA appears to be a promising solution to the problems of high spectral and energy efficiencies for 5G and beyond, given its provable benefits over OMA [15,16].

In this work, we address the problem of joint power allocation and channel assignment in underlay MD2D communications when the receivers can exploit NOMA, a channel/RB can be used by multiple MD2D groups and a MD2D group needs to simultaneously transmit over several channels to meet its rate constraint. Although the use of NOMA and SIC receivers has already been analyzed in several works [17,18] for unicast D2D communications, this paper considers the performance evaluation of NOMA for MD2D, so it generalizes previous assumptions. We formulate the optimization problem of energy efficiency (EE) under throughput constraints either for network EE or for max-min fair (MMF) EE maximization, since aggregate network performance does not capture adequately the notion of individual fairness. To that end, we extend our prior mathematical framework in [4,19] and explicitly model the existence of SIC receivers in the system. In addition, the degree of resource sharing can be adapted by design through two simple parameters: the maximum number of MD2D groups per RB (the reuse degree) and the maximum number of RBs assigned to a MD2D group (the split factor). This resource allocation problem under NOMA and SIC is solved with a two-stage approach in a semi-distributed form, much as in [4], by a combination of mathematical optimization (fractional programming) and game theory. Our results quantify the improvements gains associated to NOMA and SIC in MD2D communications, showing that NOMA is able to attain higher sum-rate, global energy efficiency and max-min rate (all simultaneously) than orthogonal transmission modes, even though it does not systematically achieve better max-min energy efficiency. Moreover, the benefits of NOMA start to appear only when there is enough contention for the shared channels and are clearer if the degrees of freedom in using the shared channels are larger.

The rest of the paper is organized as follows. Section 2 discusses the relevant literature in the context of Device-to-Device (D2D) pair/groups, MD2D and NOMA. Section 3 illustrates the system model. Then, optimal power allocation is detailed in Section 4, followed by the description of both the centralized and the distributed channel assignment solutions. Numerical results for the proposed schemes are given in Section 5. Section 6 contains some discussion and considerations on the role of NOMA in MD2D systems. Finally, concluding remarks are presented in Section 7.

2. Related Work

Both NOMA techniques and D2D communications have the potential to significantly improve energy and spectral efficiency (EE and SE, respectively) in 5G networks and beyond without demanding any modification to the deployed infrastructure [20]. However, integrating D2D technology into 5G comes with a set of technical hurdles, mainly, the co-channel interference between D2D communications and between D2D and other CUs served by the BS [21]. In a scenario where both D2D technology and NOMA are in use, the co-channel interference caused by D2D transmissions adds new challenges to the power allocation of CU transmissions.

Joint power control and channel allocation is investigated in [17] for maximizing the sum-rate of D2D pairs in a unicast transmission mode, where D2D nodes underlay a coexistent NOMA-based cellular network. A dual-based iterative decomposition approach of the optimization problem is developed to simplify the solution and determine the optimal transmission power for a set of CUs over the channels, and next the power of D2D transmitters and their allocated channels are selected. The combination of NOMA for the cellular users and D2D has also been explored in the mobile edge context (e.g., [22]) as a strategy to offload computing tasks from the edge servers. The goal is to minimize the weighted sum of users' energy consumption and the computation delay. As another example, Wang et al. [23] incorporated NOMA-based D2D in the design of an advanced H-CRAN for 5G. Other works extend the setting to the case wherein both CUs and D2D receivers can exploit NOMA to improve system performance metrics, as in [24] (maximization of the sum rate) and [25] (minimization of the sum-power in the network). However, Zhao et al. [24] used a fixed power allocation, hence it does not realize all the potential gains, while Yoon et al. [25] relied on heuristics to solve the channel assignment problem.An efficient solution for power control and channel allocation is presented in [26,27], limited to the case in which a set of D2D pairs (D2D groups) is assigned to only one resource block. An alternative is to separate the regimes based on the aggregate interference level in the network, and use D2D with SIC only when the interference is low. This is explored in [28], but the model turns out to be an impractical combinatorial problem hard to solve that needs substantial simplifications. A mixed communication system that restricts NOMA to the D2D pairs and continues to use OMA for the transmissions between CUs and the BS is analyzed in [29] and solved used matching theory. Multicast D2D groups with NOMA are the focus of the works in [18,30], but they only consider groups with two receivers in order to simplify the analysis of the rates.

Compared with the state-of-the-art, the work presented in this paper introduces the following contributions:

1. The investigated system models in the literature assume a single/group of D2D pairs can share CUs resource blocks. In this unicast communication model, a transmitter sends data to a single receiver. However, in our system model, devices with a common interest form a group, where a single transmitter multicasts data to a set of receivers (the MD2D group). Note that the MD2D mode also includes the particular case of unicast D2D communications (when the groups just include one receiver).
2. In multicast communications, group data rate is determined according to the receiver with the poorest channel quality (CQ). Therefore, we assume that the receivers in a MD2D group as well as the CUs are able to apply SIC to the stronger interference signal. This would reduce the received interference to a minimum value leading to enhancements in MD2D communication quality.
3. We model the resource allocation sub-problem in underlay MD2D using matching theory and overlapping coalition formation game to minimize harmful mutual interference as the means to maximize energy efficiency for both the system as a whole and for individual users, metrics not considered in other related works.
4. The resource allocation approach involves two design parameters (the reuse degree and the split factor) which allow considering all the possible RB sharing variants when analyzing the system behavior.
5. The power control sub-problem is optimally solved using fractional programming. In this work, we assume that a central entity is in charge of assigning transmission power to each transmitter in the network.
6. We evaluate the performance of NOMA-based systems under a broad range of resource sharing scenarios. As performance metrics, we analyze both EE and transmission rate for the whole system (global EE and sum-rate). A potential drawback of this approach is that global performance measures do not properly capture the service fairness among users. Thus, we also formulate and analyze the MMF EE and the rate for individual users.

As stated, the sharing of resources includes the possibility that a user transmits over multiple channels and that a channel may be accessed by several transmitters. Assume that r, the *reuse degree*, represents the maximum number of transmitters per RB, and that s, the *split factor*, is the maximum number of allowed channels that a transmitter can use to increase its throughput. Both r and s are system parameters, which variations yield four distinguishable resource sharing scenarios:

Scenario 1 ($r = s = 1$): Each CU shares its communication channel with a single D2D pair/MD2D group. Similarly, each D2D pair/MD2D group uses only one channel.

Scenario 2 ($s > 1, r = 1$): A D2D pair/MD2D group takes advantage of multiple cellular channels and distributes its message over them. Here, a D2D pair/MD2D group may disperse its transmission power budget or rate among the occupied channels. However, a channel cannot support more than one D2D pair/MD2D group.

Scenario 3 ($s = 1, r > 1$): Each cellular communication channel can support up to r D2D pairs/MD2D groups, which are allowed to use one cellular channel, at most. Compared to the previous scenarios, having r devices using the same channel leads to mutual interference accumulation over the CU and each D2D pair/MD2D group per channel.

Scenario 4 ($s > 1, r > 1$): Each D2D pair/MD2D group uses up to s different cellular channels. Moreover, each channel can support r D2D pairs/MD2D groups. As in Scenario 3, the accumulated mutual interference among the channel's devices may negatively affect both type of communications and limit the benefits of D2D pairs/MD2D groups in term of spectral and energy efficiency.

For clarity purpose, we give each scenario a descriptive name: Scenario 1 (Dedicated CUs), Scenario 2 (Distributed Groups), Scenario 3 (Shared CUs) and Scenario 4 (General Case). We refer the reader to Table 1 for a summary of the related work explored in this Section using the proposed taxonomy.

Table 1. D2D pair/groups, MD2D and NOMA: State of the art.

Ref.	Scenario	Approach	Model	Problem	Objective	NOMA
[17]	Dedicated CUs	Optimization	D2D pair	PC,RA	D2D sum rate	CU
[22]	Distributed Groups	Matching Theory	D2D group	PC,RA	Energy consumption, delay	CU
[25]	Distributed Groups	Optimization	D2D pair	RA,PC	Min transmission power	D2D + CUs
[26]	Dedicated CUs	Graph Theory	D2D group	PC,RA	D2D EE	D2D
[24]	Distributed Groups	Match Theory	D2D group	RA	Network sum rate	D2D + CUs
[28]	Dedicated CUs	Optimization	D2D pair	RA + Mode Selection	System sum rate	CU
[29]	Dedicated CUs	Matching Theory	D2D group	RA	System sum rate	D2D
[30]	Dedicated CUs	Game Theory	D2D group	RA	CUs throughput	D2D
[18]	Dedicated CUs	Matching Theory	D2D group	PC,RA	Maximum users SINR	D2D
[27]	Dedicated CUs	Hungarian Algorithm	D2D pairs	PC,RA	D2D energy	D2D

3. Problem Formulation

3.1. System Model

We consider a single-cell MD2D communications scenario with $K \geq 2$ MD2D groups of users and $M \geq 2$ CUs (see Figure 1). The BS communicates with the associated CUs over M orthogonal downlink channels. We identify each channel with an active CU in the downlink, thus the set of channels/CUs is denoted as $\mathcal{C} = \{C_1, \ldots, C_M\}$. The set of receivers in group k is denoted by $\mathcal{D}_k$, so $\mathcal{D} = \{\mathcal{D}_1, \ldots, \mathcal{D}_K\}$ is the set of K MD2D groups. Any MD2D group, to increase its throughput, can simultaneously transmit over multiple channels up to a maximum of s channels, where s is the *split factor*; additionally, multiple

MD2D groups can share the access to the same channel, up to a maximum of r groups per channel, where r is the *reuse degree* [4].

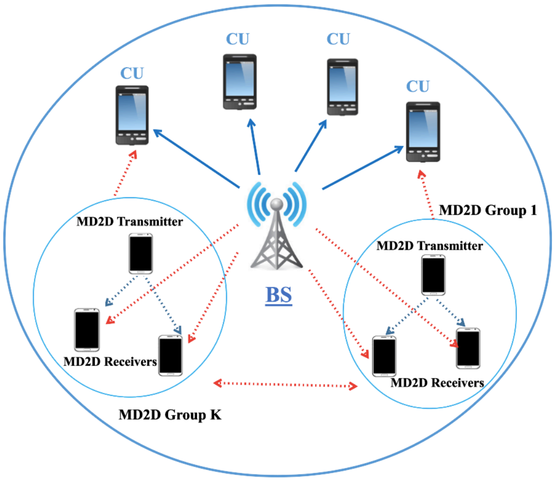

Figure 1. System model.

In this setting, the receivers in a given MD2D group have to cope with two kinds of interference:

- The CU interference is caused by the BS transmissions to the CU(s) over the channel(s) used by the MD2D group at the same time.
- The inter-groups interference is caused by the transmitters on those MD2D groups that are reusing the channel(s) the MD2D group is also accessing.

Interference cancellation of superposed signals in the same time–frequency resource is possible through SIC, where at least one user is forced to fully decode the messages of the other co-channel users, subtract them from the received signal, and decode its own message afterwards. In information theory, receiver-side SIC and superposition coding (NOMA) are known to achieve the capacity region of the Single-Input Single-Output Gaussian Broadcast Channel (SISO-BC), which is strictly larger than the capacity region achieved by orthogonal transmissions [31]. This increase in the rate is attained through increased complexity in the SIC receivers, since a subset of these have to fully recover messages directed to other users, as illustrated in Figure 2.

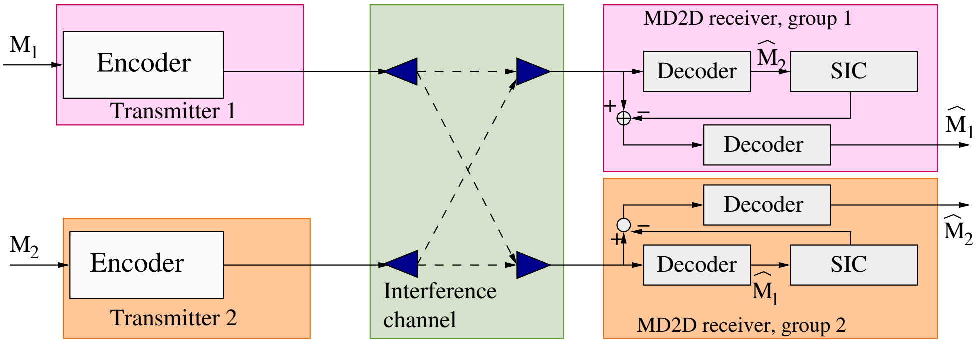

Figure 2. Communications architecture for the NOMA-SIC MD2D network.

3.2. Channel Model

The *signal to interference and noise ratio* (SINR) observed by each CU user C_i is

$$\Gamma_i = \frac{h_i p_i}{N_0 + \sum_{k \in \mathcal{D}} c_{k,i} h_{k,i} p_{k,i}}, \quad \forall C_i \in \mathcal{C}, \tag{1}$$

where h_i is the channel gain between user C_i and the BS, $h_{k,i}$ is the link gain between the transmitter in MD2D group $\mathcal{D}_k$ and the BS on channel i, p_i (respectively, $p_{k,i}$) is the transmission power used for user C_i (by the transmitter in $\mathcal{D}_k$ on channel i), N_0 is the noise density and

$$c_{k,i} = \begin{cases} b_{k,i}, & \text{if } |h_{k,i} p_{k,i}| < |h_i p_i|, \\ 0, & \text{otherwise}, \end{cases} \tag{2}$$

where $b_{k,i} = 1$ if the MD2D group $\mathcal{D}_k$ uses channel i (and 0 otherwise). Note that power-domain NOMA [6,7] is applied in (2), since the interference caused by stronger interferers—those having $|h_{k,i} p_{k,i}| > |h_i p_i|$—is suppressed. Recall that, with SIC, the receiver decodes first the stronger signals, subtracts them from the received signals, and reduces the interference to that caused by the weaker transmitters.

Similarly, the SINR observed in channel i for each receiver node $j \in \mathcal{D}_k$ is given by

$$\Gamma_{k:j,i} = \frac{g_{k:j,i} p_{k,i}}{N_0 + \tau_{k:j,i} p_i \beta_{k:j,i} + \sum_{\ell \neq k} \delta_{\ell,i} p_{\ell,i} g_{\ell:j,i}}. \tag{3}$$

In this case, the channel link gains are $g_{k:j,i}$ (the gain between transmitter k and receiver $j \in \mathcal{D}_k$ over channel i), and $\beta_{k:j,i}$, the gain between user C_i and receiver $j \in \mathcal{D}_k$. The indicator variables at the denominator are

$$\tau_{k:j,i} = \begin{cases} 1, & \text{if } |p_i \beta_{k:j,i}| < |g_{k:j,i} p_{k,i}|, \\ 0, & \text{otherwise}; \end{cases} \tag{4}$$

and

$$\delta_{\ell,i} = \begin{cases} b_{k,i}, & \text{if } |g_{\ell:j,i} p_{\ell,i}| < |g_{k:j,i} p_{k,i}|, \\ 0, & \text{otherwise}. \end{cases} \tag{5}$$

To simplify notation, we denote the total observed interference as $I_{k:j,i} = \tau_{k:j,i} p_i \beta_{k:j,i} + \sum_{\ell \neq k} \delta_{\ell,i} p_{\ell,i} g_{\ell:j,i}$, so that $\Gamma_{k:j,i} = \frac{g_{k:j,i} p_{k,i}}{N_0 + I_{k:j,i}}$.

Note that, in (2), (4) and (5), we assume perfect knowledge of channel state information (CSI). Further, the transmitters follow a distributed NOMA strategy, as shown in Figure 2, since different transmitters (the BS in each of the CUs' resource blocks or transmitters in other MD2D groups) encode their signals independently but with a common codebook. For the receivers, (2), (4) and (5) imply that those in different groups can use different SIC orderings for decoding, since the channel qualities vary.

3.3. Resource Allocation Problem

Our objective in this work is to maximize the energy efficiency of both the MD2D groups and the CUs under the constraint of a minimum rate requirement for every user. This minimum rate may be different for each group or CU, since individual groups or CU users may have diverse data rate requirements depending on the type of their applications.

The (normalized) energy efficiency, measured in bit/Hz/J, is defined for each CU user C_m as the ratio of its normalized transmission rate to the energy consumed:

$$\mathsf{EE}_{\mathsf{C}}(m) := \frac{r_m}{u_m + p_m} = \frac{\log_2(1 + \Gamma_m)}{u_m + p_m}, \quad \forall C_m \in \mathcal{C}, \tag{6}$$

where r_m denotes the transmission rate in bit/s/Hz of user C_m, u_m is the residual power consumed by user C_m when there are no data to transmit (a constant) and p_m is its transmission power. For the MD2D group $\mathcal{D}_k$, its energy efficiency is given by

$$\mathsf{EE}_{\mathsf{D}}(k) := \frac{R_k}{v_k + \sum_{m \in \mathcal{C}} P_k^{(m)}} = \frac{|\mathcal{D}_k| \sum_{m \in \mathcal{C}} \min_{j \in \mathcal{D}_k} \log_2(1 + \Gamma_{k:j,m})}{v_k + \sum_{m \in \mathcal{C}} P_k^{(m)}}, \quad \forall \mathcal{D}_k \in \mathcal{D}, \tag{7}$$

where v_k denotes the used power by the transmitter and the receivers in $\mathcal{D}_k$ group at rest and $P_k^{(m)}$ is the assigned transmission power to the transmitter in group $\mathcal{D}_k$ over channel C_m. Note that the rate R_k supportable in group k (the numerator in (7)) is limited by the receiver with the poorest channel quality, and scales with the size of the group.

The global energy efficiency (GEE) of the cellular network is the ratio between the aggregated rate and the total power needed. Thus, the optimization problem for maximizing the GEE is formulated as

$$\max_{\mathbf{p}, \mathbf{P}_k, \mathbf{B}} \frac{\sum_{m \in \mathcal{C}} \log_2(1 + \Gamma_m) + \sum_{k \in \mathcal{D}} |\mathcal{D}_k| \sum_{m \in \mathcal{C}} \min_{j \in \mathcal{D}_k} \log_2(1 + \Gamma_{k:j,m})}{\tau \sum_{m \in \mathcal{C}} p_m + \sum_{k \in \mathcal{D}} \sum_m b_{k,m} P_k^{(m)}} \tag{8}$$

$$\begin{aligned}
\text{subject to } & b_{k,m} \in \{0,1\}, \sum_{m \in \mathcal{C}} b_{k,m} \le s, \sum_{k \in \mathcal{D}} b_{k,m} \le r, && k \in \mathcal{D}, m \in \mathcal{C} && \text{(9a)} \\
& p_m \le \overline{p}, && m \in \mathcal{C} && \text{(9b)} \\
& \sum_{m \in \mathcal{C}} P_k^{(m)} \le \overline{P}_k, && k \in \mathcal{D} && \text{(9c)} \\
& \log_2(1 + \Gamma_m) \ge \underline{r}_m, && m \in \mathcal{C} && \text{(9d)} \\
& |\mathcal{D}_k| \sum_{m \in \mathcal{C}} \min_{j \in \mathcal{D}_k} \log_2(1 + \Gamma_{k:j,m}) \ge \underline{R}_k, && k \in \mathcal{D} && \text{(9e)}
\end{aligned}$$

over the variables $\mathbf{p} = (p_1, \dots, p_M)$ (the power vector allocated to the CUs), $\mathbf{P}_k = (P_k^{(1)}, \dots, P_k^{(M)})$, $k = 1, \dots, K$, (the power vector allocated to the designated transmitter of group $\mathcal{D}_k$ over the M channels) and $\mathbf{B} = [b_{k,m}]$, the $K \times M$ channel allocation matrix. The constant $\tau = \sum_{m \in \mathcal{C}} u_m + \sum_{k \in \mathcal{D}} v_k$ is just the power consumption of all the devices when there is nothing to transmit. Observe that both CU users and MD2D groups must satisfy individual average power and minimum transmission rate constraints (9b)–(9e).

A second system performance metric is the max-min fair energy efficiency (MMF-EE), which is defined as $\mathsf{EE}_{\mathsf{mmf}} = \min\{\min_m \mathsf{EE}_{\mathsf{C}}(m), \min_k \mathsf{EE}_{\mathsf{D}}(k)\}$. Clearly, the MMF-EE uniformly lower bounds the individual EE of all the users, so it provides a common quality of service to all the devices in the network. The corresponding optimization problem is, therefore, to maximize this minimum EE:

$$\max_{\mathbf{p}, \mathbf{P}_k, \mathbf{B}} \mathsf{EE}_{\mathsf{mmf}} \tag{10}$$

over the same variables as (8) and constraints (9a)–(9e).

The joint power control and resource allocation is a mixed integer non-linear problem (MINLP) which is NP-hard as proved in [19]. Therefore, we decompose it into two subproblems: (i) resource allocation; and (ii) power control. In our approach, a (sub-optimal, in general) feasible allocation of channels to MD2D groups is calculated in the first stage; then, in the second or inner stage, the optimal power for each transmitter in the system that maximizes the EE is obtained by successive convexification of the objective function and solving fractional programming problem. This process is iterated until convergence, which is guaranteed and, in our numerical tests, is reached in just a few rounds.

4. Resource Allocation: Power Control and Channel Assignment

Our schema to solve the joint power control and channel assignment problem is an alternating optimization between the discrete part and the continuous part. The latter is solved optimally given a fixed channel allocation, and, once the transmission power have been determined, the channels and users (UEs and MD2Ds) are paired. This process is repeated until convergence.

4.1. Power Control Algorithm

Suppose a fixed, static channel allocation matrix $\mathbf{B}$. Under this condition, both problems (8) and (10) seem standard convex optimizations, since (9b)–(9e) define a convex solution space, but on closer inspection it can be seen that the objective function is really not concave in $\mathbf{p}$. However, a two-step transformation suffices for converting (8) into a convex problem, and then any classical interior point method is applied to efficiently obtain its solution. Specifically, (8) is first written as $\max_{\mathbf{p}\in\mathcal{P}} f(\mathbf{p})/g(\mathbf{p})$, where $g(\mathbf{p})$ is affine. Should the numerator be concave, the well known Dinkelbach's algorithm [32] (reproduced here for completeness as Algorithm 1) could readily be used to find the solution, but concavity does not hold in this case as pointed. Thus, at every iteration of Algorithm 1 (Line 3), a concave local approximation $\tilde{f}(\cdot)$ is used in place of the true $f(\cdot)$. Specifically, let $\mathbf{p}_k(\mathbf{B})$ be the current approximation at iteration k to the optimal vector of transmission powers $\mathbf{p}^*(\mathbf{B})$ when the channels are shared as $\mathbf{B}$ dictates. Then, $f(\mathbf{p})$ is substituted by a concave $\tilde{f}(\cdot)$ such that $\tilde{f}(\mathbf{p}_k(\mathbf{B})) = f(\mathbf{p}_k(\mathbf{B}))$ and a new vector $\mathbf{p}_{k+1}(\mathbf{B})$ of power values is obtained as the maximizer point. We refer to the work in [4,33] for the proof that this procedure converges and finishes in linear time. Note that, for the optimal power vector and multiplier $(\mathbf{p}^*, \lambda^*)$, we have $f(\mathbf{p}^*) - \lambda^* g(\mathbf{p}^*) = 0$.

We conclude our discussion of the power control algorithm with the observation that, although the procedure is unchanged with respect to Hmila et al. [4], the outcome of the algorithm is *different* with OMA and NOMA for a fixed channel usage matrix $\mathbf{B}$, because the effective interference appearing in (1)–(3) is lower with NOMA, and the corresponding SINR is higher. The impact of this is discussed below when the numerical results are presented.

Algorithm 1 Optimal power control for GEE and MMF-EE (Dinkelbach's algorithm).

```
1: $\epsilon > 0, \lambda = 0, F = \infty$
2: while $F > \epsilon$ do
3:     $\mathbf{p}^\star = \arg\max_{\mathbf{p}\in\mathcal{P}}\{f(\mathbf{p}) - \lambda g(\mathbf{p})\}$
4:     $F = f(\mathbf{p}^\star) - \lambda g(\mathbf{p}^\star)$
5:     $\lambda = f(\mathbf{p}^\star)/g(\mathbf{p}^\star)$
6: end while
```

4.2. Centralized Channel Assignment: Matching Theory

We present two complementary methods to calculate the assignment of channels to users. The first resorts to matching theory and is centralized, a single entity (e.g., the BS) is in charge of selecting the best pairs CUs-MD2Ds on a given channel. Matching theory has been previously applied to this setting [34,35], but mostly in the one-to-one and one-to-many variants. In this paper, we also need to consider a many-to-many matching between CUs and MD2D groups in order to capture the *General Case* introduced in Section 2. The simpler resource sharing policies of *Dedicated CUs* and *Distributed Groups* are adequately covered by one-to-one and one-to-many matchings, respectively.

The preference function used to compute the matching is key for the accuracy of the solution. Heuristically, we use for this purpose the aggregate interference level measured at MD2D receivers and CUs. Our rationale is that mitigating the interference entails less power for a given transmission rate. In the interference-limited regime (low or medium

SINR), the rate is approximately linear in the SINR level, so we expect that the matching selected on the basis of less accumulated interference is close to optimal. Our numerical simulations confirm this observation (see also [4,19]).

4.2.1. Channel Assignment Algorithm

Formally, a channel assignment is a function between the set of MD2D groups $\mathcal{D}_k \in \mathcal{D}$ and the downlink channels, here identified with the set $\mathcal{C}$. Pairings that cause the minimum possible mutual interference are preferred, according to the following relationships.

One-to-One Matching: A one-to-one match μ is a mapping from $\mathcal{D} \cup \mathcal{C}$ to itself such that, for any $\mathcal{D}_k \in \mathcal{D}$, if $\mu(\mathcal{D}_k) \neq \mathcal{D}_k$, then $\mu(\mathcal{D}_k) \in \mathcal{C}$, and, if $\mu(C_m) \neq C_m$ for some $C_m \in \mathcal{C}$, then $\mu(C_m) \in \mathcal{D}$. The partner $\mathcal{D}_k$ is referred to as $\mu(C_m)$ if $\mu(C_m) = \mathcal{D}_k$. The preference function for setting the matching uses the received aggregate interference on each MD2D group $\mathcal{D}_k$ given by

$$\alpha_k^{(m)} = \max_{j \in \mathcal{D}_k} I_{k:j,m}, \quad \forall \mathcal{D}_k \in \mathcal{D}, \tag{11}$$

for channel m. In an analogous form, the aggregated interference seen by each CU user C_m is

$$\Gamma_m = \sum_{k \in \mathcal{D}} c_{k,m} h_{k,m} p_{k,m}, \quad \forall C_m \in \mathcal{C}. \tag{12}$$

Since Γ_m is additive, we isolate the contribution of the MD2D group $\mathcal{D}_i$ by denoting $\Gamma_m^{\mathcal{D}_i} = h_{i,m} p_{i,m}$, or, equivalently, setting $c_{i,m} = 1$ and $c_{k,m} = 0$, $\forall k \neq i$, in (12). Then, the preference relationship is defined as follows: (i) group $\mathcal{D}_k$ prefers channel C_i to C_j if $\alpha_k^{(i)} < \alpha_k^{(j)}$; and (ii) user C_m prefers group $\mathcal{D}_i$ to $\mathcal{D}_j$ if $\Gamma_m^{\mathcal{D}_i} < \Gamma_m^{\mathcal{D}_j}$. Note that, if a channel m is empty, then $\mu(C_m) = C_m$, and, when group $\mathcal{D}_k$ is forbidden to transmit on any channel, $\mu(\mathcal{D}_k) = \mathcal{D}_k$.

Many-to-One Matching: A many-to-one match μ is a mapping from $\mathcal{D} \cup \mathcal{C}$ to itself such that, for each $\mathcal{D}_k \in \mathcal{D}$, if $\mu(\mathcal{D}_k) \neq \mathcal{D}_k$, then $\mu(\mathcal{D}_k) \in \mathcal{C}$, and, if $\mu(C_m) \neq C_m$ for some $C_m \in \mathcal{C}$, then $\mu(C_m) \in \mathcal{D}$. The partner $\mathcal{D}_k$ is referred to as $\mu(C_m)$ if $\mu(C_m) = \mathcal{D}_k$. For many-to-one matches, the preference relationship is similarly defined as follows: (i) transmitter in $\mathcal{D}_k$ prefers channel C_i to channel C_j if $\alpha_k^{(i)} < \alpha_k^{(j)}$; and (ii) user C_m prefers $\mathcal{D}_i$ to $\mathcal{D}_j$ if $\Gamma_m^{\mathcal{D}_i} < \Gamma_m^{\mathcal{D}_j}$; (iii) $|\mu(C_m)| \leq r$, where r is channel m's reuse factor.

Many-to-Many Matching: For matches between arbitrary subsets of $\mathcal{D}$ and $\mathcal{C}$, the following preference relationship is defined by: (i) $\mathcal{D}_k$ prefers channel C_i to C_j if $\alpha_k^{(i)} < \alpha_k^{(j)}$; (ii) user C_m prefers $\mathcal{D}_i$ to $\mathcal{D}_j$ if $\Gamma_m^{\mathcal{D}_i} < \Gamma_m^{\mathcal{D}_j}$; (iii) $|\mu(C_m)| \leq r$, where r is channel m's reuse factor; and (iv) $|\mu(\mathcal{D}_k)| \leq s$, where s is group $\mathcal{D}_k$'s split factor.

In the centralized approach, a single controller entity first sets a preference list of the transmitters over $\mathcal{C}$ and a preference list of the channels/CUs over the K groups. If there is no sharing (this corresponds to $r = s = 1$, or dedicated CUs), then the central entity simply applies the Gale–Shapley algorithm to obtain a stable matching between channels and MD2D groups [35]. For the resource sharing cases given when $r > 1$ or $s > 1$, multiple transmitters can share a channel and multiple channels can be used by a transmitter, and a many-to-one matching or many-to-many matching is decided with Algorithm 2. Now, in contrast to the dedicated case, the decision to put MD2D group $\mathcal{D}_k$ into channel C_m depends on the co-channel inter-group interference produced by CU C_m and possibly other groups $\mathcal{D}_j, j \neq k$, already using the same RB. Likewise, the acceptance or rejection of a channel C_m for a new MD2D group may differ according to the aggregated interference (12). Thus, Algorithm 2 executes the Gale–Shapley exchange iteratively, and in each round one MD2D group is selected and assigned. To that end, the controller first calculates interference as if it would in a one-to-one matching. During the round, those groups whose rank in the channels' preference lists is not the highest are deferred until some next iteration, and the SINR are recalculated based on the current assignment to evaluate the increase on channel C_m due to the addition of some of the unmatched groups. Actually, all the values of the aggregated interference are re-evaluated, irrespective of the status of the group, matched or unmateched. Note that, through the use of (11) and (12), the matching algorithm is considering the SIC decoding for the evaluation of the interference levels. This is an important difference with the previous literature on the use of matching theory for wireless networks [18].

Algorithm 2 NOMA interference-based matching algorithm.

```
Set up preference list $D2MD_{PL}, \forall \mathcal{D}_k \in \mathcal{D}$, from (11)
Set up preference list $CUE_{PL}, \forall C_m \in \mathcal{C}$, from (12)
Initial unmatched groups list $\mathcal{L} = \mathcal{D}$
Free groups list $\mathcal{F} = \emptyset$
$r \leftarrow 1$
while $C_m$ occupancy $< r$ do
    while $\mathcal{L} \neq \emptyset$ do
        repeat
            if $C_m$ occupancy $< r$ then
                Allocate $C_m$ to group $\mathcal{D}_k$;
            else if $C_m$ is allocated to $\mathcal{D}_{k'}$ and $\mathcal{D}_k$ is preferred over $\mathcal{D}_{k'}$ then
                Reject $\mathcal{D}_{k'}$
                Keep $\mathcal{D}_k$
                $\mathcal{L} \leftarrow \mathcal{L} \cup \{\mathcal{D}_{k'}\}$
            else
                Keep $\mathcal{D}_{k'}$ and reject $\mathcal{D}_k$
            end if
            Update $|\mathcal{D}_k|$ and $|\mathcal{D}_{k'}|$
            if $|\mathcal{D}_k| = s$ then
                $\mathcal{L} \leftarrow \mathcal{L} \setminus \{\mathcal{D}_k\}$
            else if $\forall C_i \in \mathcal{C}$, $\mathcal{D}_k$ is rejected then
                $\mathcal{F} \leftarrow \mathcal{F} \cup \{\mathcal{D}_k\}$
            end if
        until all preferences of group $\mathcal{D}_k$ are tested or $|\mathcal{D}_k| = s$
    end while
    Calculate interference values (11) and (12)
    $\mathcal{L} = \mathcal{F}$
    $r \leftarrow r + 1$
end while
```

4.2.2. Stability

The matching algorithm produces a stable outcome if no two pairs of agents (c_1, d_1) and (c_2, d_2) can be found such that the global utility function increases by swapping the pairings, namely using (c_1, d_2) and (c_2, d_1) instead. If some pair with this property exists, it is called a *blocking pair*, so the partnership μ is stable if none pair $(\mathcal{D}_k, C_m)$ forms a blocking pair for any $\mathcal{D}_k \in \mathcal{D}$ and $C_m \in \mathcal{C}$ that are not currently matched to each other. This means that both group $\mathcal{D}_k$ and channel C_m prefer each other more than their current partners in the matching. To prove the stability we need to show that these two conditions cannot hold simultaneously. Assume that $\mathcal{D}_{k'}$ prefers C_m, so it sends a preference message to C_m based on its individual preferences $\mu(\cdot)$. Consequently, $\mu(\mathcal{D}_k) \neq C_m$ as $\mathcal{D}_k$ has a lower priority by C_m according to μ. This shows that, even though C_m is $\mathcal{D}_k$'s favorite partner, C_m does not have incentives toward being matched to $\mathcal{D}_k$. Thus, the first condition fails. The second condition can be proved along the same argument, and the pair $(\mathcal{D}_k, C_m)$ cannot be a blocking pair for μ, too. Therefore, the relationship and pairings chosen by the matching algorithm are stable.

4.3. Distributed Channel Assignment: Coalition Formation

In Section 4.2, we suppose that a central entity has perfect CSI and uses that for running the matching algorithms. In this section, we show that channel assignment problem can be solved almost optimally using coalitional game-theoretic approach. Consider a coalition game **G**, which is defined by the triplet $(\mathcal{N}, \mathsf{v}, \mathcal{S})$:

1. $\mathcal{N} = \mathcal{C} \cup \mathcal{D}$ is the set of players, with $\mathcal{C}$ and $\mathcal{D}$ denoting the sets of CUs and MD2D groups.
2. v is the valuation function that gives the value of a coalition in a game. This is a set function that maps each $\mathcal{S}_i \subseteq \mathcal{N}$ to real non-negative number interpreted as the absolute value of the coalition.
3. $\mathcal{S} = \{\mathcal{S}_1, \mathcal{S}_2, \ldots, \mathcal{S}_n\}$ is the set of formed coalitions namely the coalition structure. Here, each coalition $\mathcal{S}_i$ is a subset of $\mathcal{N}$ ($\mathcal{S}_i \subseteq \mathcal{N}$, for $i = 1, \ldots, n$).

In particular, the definition does not imply that two coalitions are disjoint or that all the agents (CUs and MD2D groups) are part of a coalition. Without loss of generality, every singleton can be regarded as a coalition itself. We further recall some basic notions in coalitional game theory, which are useful below for our discussion. First, agents can have transferable utilities if they are allowed to transfer in a lossless way part of their individual utility to other agents. A transferable utility implies that the coalition structure determines completely the value of the coalition, and this value is totally independent of how the remaining players cooperate. A second important consequence for our purposes is that this class of coalition games have a non-superadditive valuation function. In turn, one can conclude from this that the game will never end up forming a grand coalition containing all the players, because a partition would attain a higher aggregated value. In our case, a grand coalition would lead to a high level of interference which would limit the rates. Hence, using interference for payoffs, utilities and valuations provides the correct signals to promote more efficient solutions.

The proposed coalition game applies repeatedly, in an asynchronous and distributed manner, the following two rules for merging new coalitions or dividing existing ones:

Merge coalitions Any subset of coalitions $\{\mathcal{S}_1, \ldots, \mathcal{S}_l\}$ may be merged whenever the merged form is preferred by the players. The preference relationship is the same defined in Section 4.2.1 for the centralized solution approach.

Split coalitions Any coalition $\bigcup_{j=1}^{l} \mathcal{S}_j$ may be split whenever the split form is preferred by the players.

The principle that lies under the rule for coalition formation is that some MD2D group $\mathcal{D}_k$ can depart from coalition $\mathcal{S}_i$ and become a new member of coalition $\mathcal{S}_j$ only if these conditions are met:

1. The weakest receiver in group $\mathcal{D}_k$ suffers less individual interference if the group is moved to $\mathcal{S}_j$ (recall the the receiver with the poorest channel sets the transmission rate for the group).
2. The total mutual interference of coalition $\mathcal{S}_j \cup \mathcal{D}_k$ does not increase: $\mathsf{v}(\mathcal{S}_j \cup \mathcal{D}_k) \leq \mathsf{v}(\mathcal{S}_j)$. This condition is essential for guaranteeing convergence of the algorithm. It simply states that movements among coalitions are allowed only if they improve the local value of the coalition.
3. The new coalition structure $\mathcal{S}'$ results in less total interference than the current one: $\mathsf{v}(\mathcal{S}') \leq \mathsf{v}(\mathcal{S}) \Rightarrow \mathcal{S}'$ is preferred to $\mathcal{S}$. In other words, only movements that also improve the total value of the network are approved.

The splitting of some formed coalition follows exactly the same three conditions, but reversing the direction (or the indices).

In Algorithm 3, we initially assume that each coalition has either one CU or one MD2D group with no other members. Then, these individual players take actions i.e, to merge and/or split until M coalitions emerge where each coalition has a single CUs and multiple MD2D. According to the constraints (see Algorithm 3), every coalition can host a maximum of r MD2D groups, and conversely a maximum of s coalitions have a given group as one of its members. In view of the conditions for the formation of cooperative coalitions, their members will likely be a subset of users whose co-channel interference (both intra- and inter-group) is a local minimum. Although minimizing interference is not equivalent to maximizing energy efficiency, two reasons support this choice. Firstly, it is clear that a lower interference level from transmitters in other groups will lead to use potentially less transmission power in a given group sharing the same channel(s) (for a given target rate). Keeping the rate constant (or increasing it) with less consumption of energy gives obviously improved EE. Note that this is basically a greedy argument, since the EE is increased locally, but it is plausible that the global EE also improves in a majority of the network configurations. Secondly, the minimum transmission rates are more easily achievable when the interference level is kept under tight control. This implies that the outcome of the coalition game is less sensitive and more robust to imperfect or delayed CSI. Algorithm 3 always converges to a stable set of coalitions. Here, a stable coalition is defined as one in which the coalition structure cannot be further changed by the merge or split actions presented above, and such that it maximizes the sum of utilities for the players. We refer the readers to the work of Hmila et al. [4] for a proof of this result.

4.4. *Algorithmic Complexity*

The resource allocation sub-problem algorithms, both the semi-distributed and the centralized, terminate after at most r rounds, where r is the reuse degree. In each round, the semi-distributed algorithm performs a merge and/or split move, while the centralized accepts or rejects a new member. This is repeated for at most M channels for all the coalitions or the free groups, so the complexity is no larger than $O(MK^2)$. The power control sub-problem is solved through the Dinkelbach's algorithm, which has a sublinear complexity, i.e., superlinear convergence rate [36], and the evaluation of the EE functions takes constant time. The number of variables/constraints is $O(M + K)$ (rate and power for each user). Therefore, the overall worst-case complexity is $O(rMK^2(M + K))$.

Algorithm 3 NOMA based merge-and-split for the coalitional game.

```
$r_t \leftarrow 1; \mathcal{S}_m \leftarrow \{\}, m = 1, \dots, M$
repeat
    while $r_t \leq r$ do
        Identify a feasible $p_k^{(m)}, p_m, k = 1, \dots, K, m = 1, \dots, M$.
        Set $R_k = \sum_{m \in \mathcal{C}} \log_2(1 + \Gamma_m) + \sum_{k \in \mathcal{D}} |\mathcal{D}_k| \sum_{m \in \mathcal{C}} \min_{j \in \mathcal{D}_k} \log_2(1 + \Gamma_{k:j,m})$
        Set $r_m = \log(1 + \Gamma_m), m = 1, \dots, M$
        Sort $\mathcal{S}_m$ in ascending order using (11)
        repeat
            for $m = 1, \dots, M$ do
                Choose $\mathcal{D}_k \in \mathcal{S}_m$
                if $\mathsf{v}(\mathcal{S}_m \setminus \mathcal{D}_k) < \mathsf{v}(\mathcal{S}_m \setminus \mathcal{D}_{k'})$ for some $\mathcal{D}_{k'}$ then
                    $\mathcal{S}_m \leftarrow \mathcal{S}_m \setminus \mathcal{D}_k$
                end if
                Choose $\mathcal{D}_k \notin \mathcal{S}_m$
                if $\mathsf{v}(\mathcal{S}_m \cup \mathcal{D}_k) < \mathsf{v}(\mathcal{S}_m \cup \mathcal{D}_{k'})$ for some $\mathcal{D}_{k'}$ then
                    $\mathcal{S}_m \leftarrow \mathcal{S}_m \cup \mathcal{D}_k$
                end if
            end for
            Update aggregate interference values with (11) and (12)
        until all $\mathcal{D}_k \in \mathcal{S}_m$ are tried or $|\{\mathcal{D}_k : \mathcal{D}_k \in \mathcal{S}_m\}| = s$
        $r_t \leftarrow r_t + 1$
    end while
until $\mathcal{S} = \{\mathcal{S}_1, \dots, \mathcal{S}_M\}$ is a stable coalition structure
Optimize $p_k^{(m)}, p_m, k = 1, \dots, K, m = 1, \dots, M$, using Algorithm 1
```

5. Numerical Results

In this section, we numerically evaluate the performance of both the centralized and the semi-distributed resource allocation approaches when NOMA is involved using MATLAB and the CVX mathematical optimization package.

We simulated a cell of radius equal to 500 m with $\lambda = 250$ users (CUs and D2D users). Users are spatially distributed following a standard homogeneous Poisson point process (PPP) [37,38]. Different from other works in the literature, we used two different clustering techniques to form the groups: K-Nearest Neighbor (KNN) and Distance Limit (DL) algorithms. Initially, head clusters (the transmitters per group) are randomly selected with both techniques. However, whereas KNN permits to form homogeneous (equal sized) groups, DL defines the group area as a disk around the group transmitter, so it allows us to consider groups of different sizes, including unicast transmissions as a special case. Based on previous results [39], we remark that the clustering technique (i.e., using KNN or DL) usually has little impact on the individual or aggregate behavior [39]. Therefore, we mostly present our results for only one of the algorithms, with similar conclusions being valid for the other one. For resource sharing, CUs having the best channel qualities are selected to share their RBs with MD2D groups. The received signals are assumed to weaken with path loss according to $P_r = P_t(1 + (d/d_0)^{\alpha})$ where P_r is the received power, P_t is the transmitted power, d_0 is a reference distance (100 m in our case) and $\alpha \geq 2$ is the path loss exponent. The rest of physical system parameters are summarized in Table 2.

Table 2. Simulation parameters.

Parameter	Value
Cell radius	500 m
Reuse factor (r)	$\{2, 3\}$
Network density	250 devices/cell
Split factor (s)	$\{3, 4\}$
Path loss exponent	2.5
Minimum transmission rate	$\{0.1, 0.5\}$ bit/s/Hz
Number of CU users (M)	$\{5, 6, 8\}$
Maximum transmission powers	$[-5, 25]$ dBm
Number of MD2D groups (K)	$\{4, 10, 15\}$
Number of receivers $\mathcal{D}_j$	$\{3, 4, 5\}$
Circuit power	10 dBm

5.1. Distributed Resource Allocation

Figure 3 illustrates the impact of NOMA in the semi-distributed (coalition-based) approach for the GEE and the aggregated rate. The spatial configuration used $K = 4$ and $K = 5$ groups, a number of CUs between 2 and 5, and a minimum rate of 0.1 bit/s/Hz. As shown, NOMA improves the GEE within the various resource sharing settings. Similarly, the aggregated rate significantly increases with NOMA (when compared to treating interference as noise), around a 30% in the general sharing case where the reuse of channels implies that receivers can exploit SIC for decoding weak signals. Note that, since the GEE remains almost constant, this means that the higher consumed energy is effectively converted into communication rate, as is expected with NOMA.

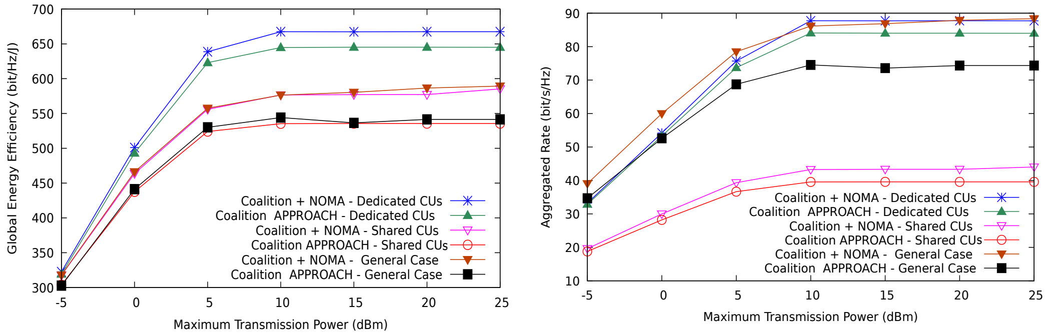

Figure 3. Global energy efficiency and aggregated rate using KNN clustering.

The minimum rate and EE_{mmf} are depicted in Figure 4 for the coalition-based approach when NOMA is used in a shared CUs setting with $r = 2$ and $r = 3$. This time, the clusters are formed using DL, with the average number of receivers set to 3. NOMA helps to increase EE_{mmf} and the transmission rate simultaneously, but the gain is clearly larger if the reuse factor r grows, since this allows for more degrees of freedom when sharing the channels. Therefore, NOMA can only realize its full potential when the co-channel interference level prevents the system to attain the target transmission rates, while it just provides a marginal gain in the high SNR regime. As a consequence, NOMA is mostly useful in dense wireless systems when the number of active users is high, even if the transmission rate required by the devices is not particularly high (e.g., mMTC in 5G).

For the general sharing scenario ($r = 2$, $s = 3, 4$), Figure 5 shows that using NOMA is not detrimental to the EE_{mmf} or the worst rate for any receiver. We used 6–8 CUs and 4 MD2D groups in both simulation cases. Further, note that the minimum rate per channel was set to 0.5 bit/s/Hz. Thus, although using NOMA leads the network to consume more power for transmitting, this is again not wasted; it is instead used to achieve the spatial degrees of freedom of the multi-user channels. Figure 6 shows precisely the increase in the consumed power between the partial sharing (shared CUs) and the full sharing (general) cases, confirming that aggregate interference is not only well controlled by the allocation algorithm (right), but also used to extract information under receiver SIC.

The impact of NOMA is also positive as the group sizes increase in the general sharing scenario (Figure 7). Both the GEE and the achieved sum-rate improve substantially with more receivers per group, and the increase is faster in the low SNR region. We emphasize that the results are shown for an *average* number of receivers per group, since the device locations and the cluster formation are random. With the parameters listed in Table 2, the maximum distance between the receiver and the transmitter happens to be within 50–80 m.

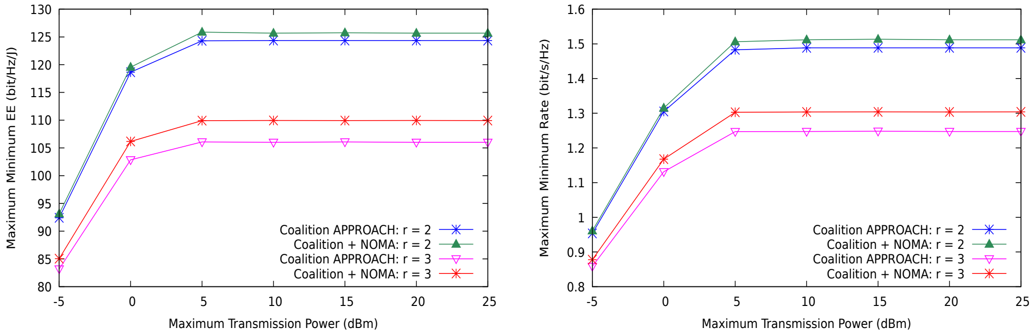

Figure 4. EE_{mmf} and max-min rate for the shared CUs case using DL clustering.

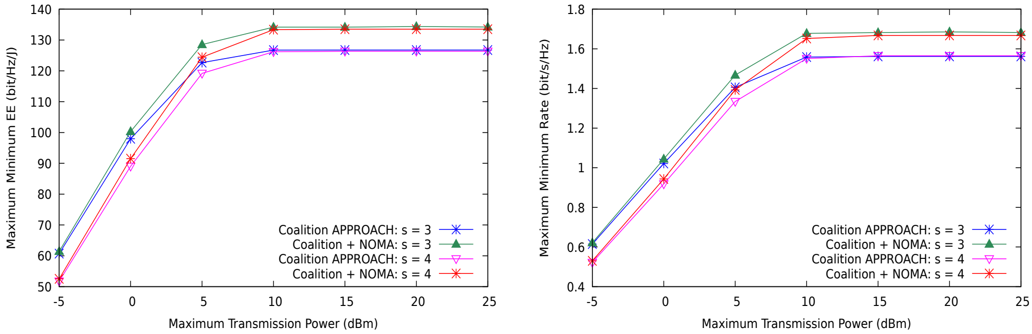

Figure 5. EE_{mmf} and max-min rate for the general case using DL clustering.

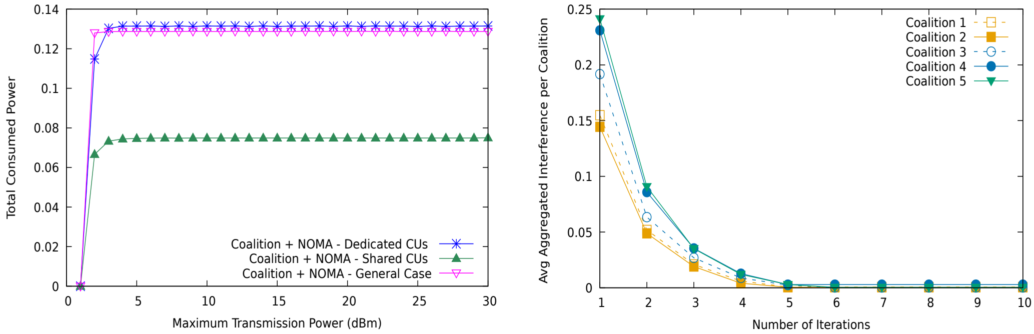

Figure 6. Total consumed power and interference level per coalition with NOMA and the distributed coalition game.

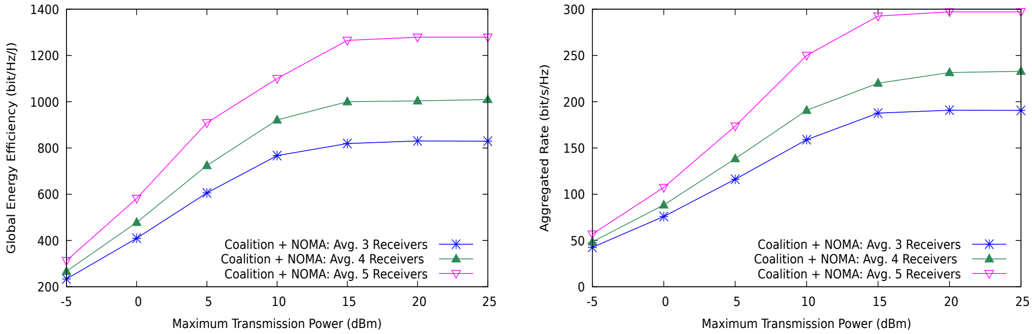

Figure 7. Global energy efficiency and aggregated rate with NOMA and the distributed coalition game for the general case with different group sizes.

5.2. Comparison to Optimal Resource Allocation

We now analyze the performance of NOMA compared to the optimal centralized resource allocation solution. Although the centralized approach is not generally applicable in large networks, its performance results set a baseline to assess the achievable gains with NOMA and SIC with respect to the system parameters (i.e., the number of groups and receivers per group).

First, we show in Figure 8 that NOMA is able to simultaneously attain better global energy efficiency and higher sum-rate than OMA in the network. This implies that NOMA better approaches the capacity region of the MD2D cell with similar or lower consumption of energy, thus demonstrating that this communication strategy clearly outperforms OMA in this scenario. However, the optimality of NOMA is still unclear since we do not currently know the exact capacity region of the system for arbitrary sizes. Accordingly, the plots in Figure 8 must be considered as an achievable inner approximation, yet with a gap to the theoretically optimal system performance. We discuss below whether NOMA alone can reasonably be the best encoding and decoding strategy for optimizing multi-user communication systems. The gain over OMA is also notable in the case of shared CUs, or partial sharing, in both of the metrics (see Figure 9). In addition, observe that the difference between NOMA and OMA (referred to as the matching approach) increases with r, the reuse factor.

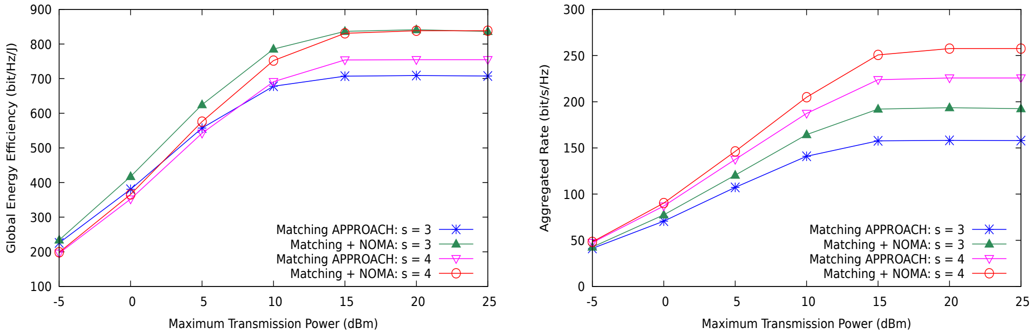

Figure 8. Global energy efficiency and aggregated rate with the matching theory based solution vs. NOMA for the general case.

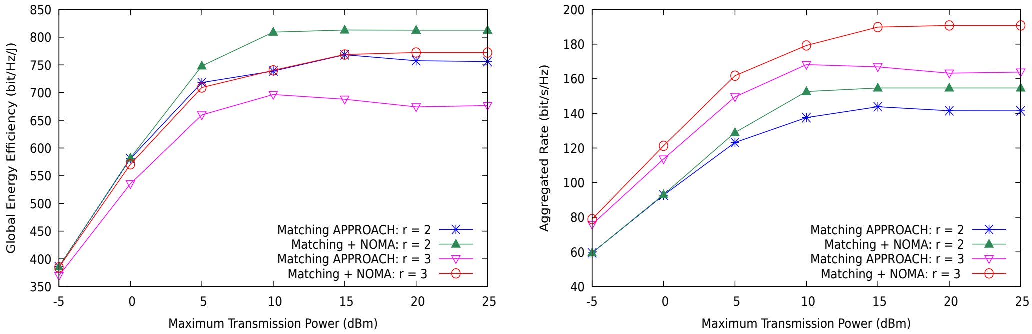

Figure 9. Global energy efficiency and aggregated rate with the matching theory based solution vs. NOMA for the shared CUs case.

Nevertheless, for the max-min performance metrics, the conclusions are slightly different. As shown in Figure 10, $\mathsf{EE}_{\mathsf{mmf}}$ is actually worse with NOMA than with OMA for a given number of channel sharing degrees of freedom (only the cases of $r = 2$ and $r = 3$ are shown). We recall that NOMA can use more power in the transmitters, since part of the interference is removed by SIC processing at the receivers, but, even under perfect SIC and strict power control, the receivers with stronger channels can experience lower $\mathsf{E}_{\mathsf{mmf}}$. This is not as unfavorable as it appears, because Figure 10 (right) shows that the max-min rate actually increases with NOMA.

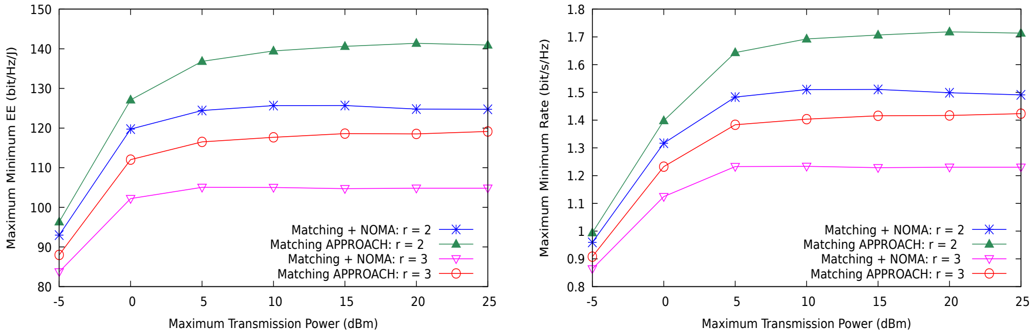

Figure 10. $\mathsf{EE}_{\mathsf{mmf}}$ and max-min rate with the proposed matching theory based solution vs. NOMA for the shared CUs case.

6. Discussion

The results presented in the previous section show that, when channel reuse is allowed in MD2D communications, NOMA and SIC-enabled receivers can attain higher sum-rate than OMA in a consistent manner, i.e., almost independently of the number of groups and the number of receivers per group. This is possible at the same energy efficiency for the network, but not generally for the MMF energy efficiency, which can be worse in some cases. However, the max-min rate is still improved with NOMA, although at the cost of consuming more energy.

In modern wireless communication systems, transmitters and receivers come equipped with multiple antennas. Our model only considers single-antenna sources and receivers, so it is natural to investigate the role of NOMA and MIMO together, since it is well known that NOMA is optimal from an information theoretical point for the single-user broadcast channel. Given that MIMO and NOMA are far better than point-to-point and OMA systems, respectively, it seems plausible that their combination yields even better benefits, at least in the sum multiplexing rate. Contrary to intuition, some recent works [40,41] unveil several shortcomings and misconceptions about multi-antenna NOMA and show that linear precoding (for the downlink) and/or rate-splitting can be strictly better than NOMA in many common network settings. Moreover, both rate splitting and precoding require less complex receivers, so their implementation is easier. In view of this, more research on the contexts where NOMA is preferable over the latter strategies is necessary to determine under which conditions a receiver-only SIC approach is optimal in terms of sum-rate and multiplexing gains. In the special case of MD2D communications, one could consider an adaptive hybrid system where NOMA and rate-splitting are used depending on the number of receivers in the D2D group, for instance, due to the similar complexity of both techniques for a small number of receivers.

7. Conclusions

In this paper, we consider the inclusion of NOMA into an optimization framework for designing multicast D2D communication systems. We show first that the base mathematical framework can be adapted without much difficulty to the non-orthogonal multiple access case for the downlink part, and that the centralized and distributed optimization procedures only need some technical changes to work properly. Next, we conducted a set of numerical experiments to evaluate the performance gains achievable with NOMA in our context, focused on global and max-min energy efficiency, and global and max-min sum rate for the receivers. Our results suggest that NOMA is efficient for the sum rate (in both cases) and for the global energy efficiency, but in contrast it may not be efficient for the max-min energy efficiency. Moreover, our results were obtained under QoS constraints for the rate, thus MD2D with NOMA can cope with high system-level throughput, reliability and heterogeneity in 5G wireless networks. While our focus was on the downlink communication, the same model and techniques can be used for the uplink direction, where the BS applies SIC for separating the signals. Nevertheless, this is a single-receiver model which has already received much attention in the literature.

While the results in the paper highlight the performance improvements that NOMA can offer for reusing the transmission channels among disjoint groups of co-located (or nearby) transmitters–receivers, there are still many issues worth investigating before fully understanding the interplay between NOMA and D2D, such as the role (architectural and related to performance) of NOMA in enhanced mMTC and eMBB services, massive IoT with short-packet communications and D2D networks. Performance analysis of NOMA in D2D and MD2D with full-duplex nodes is also a promising research direction.

Author Contributions: M.F.-V. and M.H. contributed to the conceptualization and methodology developed in this work; all authors contributed equally to the writing, original draft preparation and review; M.H. developed the simulation software and conducted the numerical experiments; M.R.-P. and S.H.-A. executed the validation of results; and general supervision of the work was done by M.F.-V. These authors contributed equally to this work. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the Spanish Ministry of Science under grant numbers PID2020-113795RB-C33 (COMPROMISE), PID2020-113240RB-I00 (ICARUS) and RED2018-102585-T (GO2EDGE). These projects receive financial support from the UE's ERDF funds.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

Conflicts of Interest: The authors declare no conflict of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript, or in the decision to publish the results.

References

1. Agiwal, M.; Roy, A.; Saxena, N. Next generation 5G wireless networks: A comprehensive survey. *IEEE Commun. Surv. Tutor.* **2016**, *18*, 1617–1655. [CrossRef]
2. Meshgi, H.; Zhao, D.; Zheng, R. Optimal Resource Allocation in Multicast Device-to-Device Communications Underlaying LTE Networks. *IEEE Trans. Veh. Technol.* **2017**, *66*, 8357–8371. [CrossRef]
3. Ye, Q.; Al-Shalash, M.; Caramanis, C.; Andrews, J.G. Distributed Resource Allocation in Device-to-Device Enhanced Cellular Networks. *IEEE Trans. Commun.* **2015**, *63*, 441–454. [CrossRef]

4. Hmila, M.; Fernandez-Veiga, M.; Perez, M.R.; Herreria-Alonso, S. Distributed Energy Efficient Channel Allocation in Underlay Multicast D2D Communications. *IEEE Trans. Mob. Comput.* **2020**. [CrossRef]
5. Elnourani, M.; Deshmukh, S.; Beferull-Lozano, B. Distributed Resource Allocation in Underlay Multicast D2D Communications. *IEEE Trans. Commun.* **2021**. [CrossRef]
6. Dai, L.; Wang, B.; Yuan, Y.; Han, S.; Chih-lin, I.; Wang, Z. Non-orthogonal multiple access for 5G: solutions, challenges, opportunities, and future research trends. *IEEE Commun. Mag.* **2015**, *53*, 74–81. [CrossRef]
7. Ding, Z.; Liu, Y.; Choi, J.; Sun, Q.; Elkashlan, M.; Chih-Lin, I.; Poor, H.V. Application of Non-Orthogonal Multiple Access in LTE and 5G Networks. *IEEE Commun. Mag.* **2017**, *55*, 185–191. [CrossRef]
8. Cover, T. Broadcast channels. *IEEE Trans. Inf. Theory* **1972**, *18*, 2–14. [CrossRef]
9. Li, J.; Huang, S. Delay-Aware Power Control for D2D Communication With Successive Interference Cancellation and Hybrid Energy Source. *IEEE Wirel. Commun. Lett.* **2017**, *6*, 806–809. [CrossRef]
10. Kazmi, S.M.A.; Tran, N.H.; Ho, T.M.; Manzoor, A.; Niyato, D.; Hong, C.S. Coordinated Device-to-Device Communication With Non-Orthogonal Multiple Access in Future Wireless Cellular Networks. *IEEE Access* **2018**, *6*, 39860–39875. [CrossRef]
11. Ali, M.S.; Tabassum, H.; Hossain, E. Dynamic User Clustering and Power Allocation for Uplink and Downlink Non-Orthogonal Multiple Access (NOMA) Systems. *IEEE Access* **2016**, *4*, 6325–6343. [CrossRef]
12. Zhao, J.; Liu, Y.; Chai, K.K.; Chen, Y.; Elkashlan, M. Joint Subchannel and Power Allocation for NOMA Enhanced D2D Communications. *IEEE Trans. Commun.* **2017**, *65*, 5081–5094. [CrossRef]
13. Liu, Y.; Qin, Z.; Elkashlan, M.; Ding, Z.; Nallanathan, A.; Hanzo, L. Nonorthogonal Multiple Access for 5G and Beyond. *Proc. IEEE* **2017**, *105*, 2347–2381. [CrossRef]
14. Shin, W.; Vaezi, M.; Lee, B.; Love, D.J.; Lee, J.; Poor, H.V. Non-Orthogonal Multiple Access in Multi-Cell Networks: Theory, Performance, and Practical Challenges. *IEEE Commun. Mag.* **2017**, *55*, 176–183. [CrossRef]
15. Zeng, M.; Yadav, A.; Dobre, O.A.; Tsiropoulos, G.I.; Poor, H.V. Capacity Comparison Between MIMO-NOMA and MIMO-OMA With Multiple Users in a Cluster. *IEEE J. Sel. Areas Commun.* **2017**, *35*, 2413–2424. [CrossRef]
16. Wei, Z.; Yang, L.; Ng, D.W.K.; Yuan, J.; Hanzo, L. On the Performance Gain of NOMA Over OMA in Uplink Communication Systems. *IEEE Trans. Commun.* **2020**, *68*, 536–568. [CrossRef]
17. Pan, Y.; Pan, C.; Yang, Z.; Chen, M. Resource allocation for D2D communications underlaying a NOMA-based cellular network. *IEEE Wirel. Commun. Lett.* **2017**, *7*, 130–133. [CrossRef]
18. Baidas, M.W.; Bahbahani, M.S.; Alsusa, E.; Hamdi, K.A.; Ding, Z. Joint D2D Group Association and Channel Assignment in Uplink Multi-Cell NOMA Networks: A Matching-Theoretic Approach. *IEEE Trans. Commun.* **2019**, *67*, 8771–8785. [CrossRef]
19. Hmila, M.; Fernández Veiga, M.; Rodríguez Pérez, M.; Herrería Alonso, S. Energy Efficient Power and Channel Allocation in Underlay Device to Multi Device Communications. *IEEE Trans. Commun.* **2019**, *67*, 5817–5832. [CrossRef]
20. Ding, Z.; Lei, X.; Karagiannidis, G.K.; Schober, R.; Yuan, J.; Bhargava, V.K. A survey on non-orthogonal multiple access for 5G networks: Research challenges and future trends. *IEEE J. Sel. Areas Commun.* **2017**, *35*, 2181–2195. [CrossRef]
21. Denis, J.; Assaad, M. Interference Management in D2D. In *Wiley 5G Ref: The Essential 5G Reference Online*; Wiley: New York, NY, USA; 2019; pp. 1–20.
22. Diao, X.; Zheng, J.; Wu, Y.; Cai, Y. Joint computing resource, power, and channel allocations for D2D-assisted and NOMA-based mobile edge computing. *IEEE Access* **2019**, *7*, 9243–9257. [CrossRef]
23. Wang, J.; Song, X.; Dong, L. Resource Allocation and EE-SE Tradeoff for H-CRAN with NOMA-Based D2D Communications. *KSII Trans. Internet Inf. Syst.* **2020**, *14*. [CrossRef]
24. Zhao, J.; Liu, Y.; Chai, K.K.; Chen, Y.; Elkashlan, M.; Alonso-Zarate, J. NOMA-based D2D communications: Towards 5G. In Proceedings of the 2016 IEEE Global Communications Conference (GLOBECOM), Washington, DC, USA, 4–8 December 2016, pp. 1–6.
25. Yoon, T.; Nguyen, T.H.; Nguyen, X.T.; Yoo, D.; Jang, B. Resource allocation for NOMA-based D2D systems coexisting with cellular networks. *IEEE Access* **2018**, *6*, 66293–66304. [CrossRef]
26. Alemaishat, S.; Saraereh, O.A.; Khan, I.; Choi, B.J. An efficient resource allocation algorithm for D2D communications based on NOMA. *IEEE Access* **2019**, *7*, 120238–120247. [CrossRef]
27. Al-kahtani, M.S.; Ferdouse, L.; Karim, L. Energy Efficient Power Domain Non-Orthogonal Multiple Access Based Cellular Device-to-Device Communication for 5G Networks. *Electronics* **2020**, *9*, 237. [CrossRef]
28. Dai, Y.; Sheng, M.; Liu, J.; Cheng, N.; Shen, X.; Yang, Q. Joint mode selection and resource allocation for D2D-enabled NOMA cellular networks. *IEEE Trans. Veh. Technol.* **2019**, *68*, 6721–6733. [CrossRef]
29. Budhiraja, I.; Kumar, N.; Tyagi, S. Cross-Layer Interference Management Scheme for D2D Mobile Users Using NOMA. *IEEE Syst. J.* **2020**. [CrossRef]
30. Gu, W.; Zhu, Q. Stackelberg Game Based Social-Aware Resource Allocation for NOMA Enhanced D2D Communications. *Electronics* **2019**, *8*, 1360. [CrossRef]
31. Tse, D.; Viswanath, P. *Fundamentals of Wireless Communication*; Cambridge University Press: Cambridge, UK, 2005. [CrossRef]
32. Dinkelbach, W. On Nonlinear Fractional Programming. *Manag. Sci.* **1967**, *13*, 492–498. [CrossRef]
33. Zappone, A.; Bjornson, E.; Sanguinetti, L.; Jorswieck, E. Globally Optimal Energy-Efficient Power Control and Receiver Design in Wireless Networks. *IEEE Trans. Signal Process.* **2017**, *65*, 2844–2859. [CrossRef]

34. Gu, Y.; Zhang, Y.; Pan, M.; Han, Z. Matching and cheating in device to device communications underlying cellular networks. *IEEE J. Sel. Areas Commun.* **2015**, *33*, 2156–2166. [CrossRef]
35. Bayat, S.; Li, Y.; Song, L.; Han, Z. Matching theory: Applications in wireless communications. *IEEE Signal Process. Mag.* **2016**, *33*, 103–122. [CrossRef]
36. Luo, F.L. *Signal Processing for 5G*; John Wiley and Sons Ltd.: Hoboken, NJ, USA, 2016.
37. Ak, S.; Inaltekin, H.; Poor, H.V. Gaussian approximation for the downlink interference in heterogeneous cellular networks. In Proceedings of the 2016 IEEE International Symposium on Information Theory (ISIT), Barcelona, Spain, 10–15 July 2016; pp. 1611–1615. [CrossRef]
38. Ak, S.; Inaltekin, H.; Poor, H.V. A Tractable Framework for the Analysis of Dense Heterogeneous Cellular Networks. *IEEE Trans. Commun.* **2018**, *66*, 3151–3171. [CrossRef]
39. Hmila, M.; Fernández-Veiga, M. Energy-Efficient Power Control and Clustering in Underlay Device to Multi-device Communications. In *Lecture Notes in Computer Science*; Springer Nature Switzerland AG.: Cham, Switzerland 2017; pp. 194–206. [CrossRef]
40. Senel, K.; Cheng, H.V.; Bjornson, E.; Larsson, e.g., What Role can NOMA Play in Massive MIMO? *IEEE J. Sel. Top. Signal Process.* **2019**, *13*, 597–611. [CrossRef]
41. Clerckx, B.; Mao, Y.; Schober, R.; Jorswieck, E.; Love, D.J.; Yuan, J.; Hanzo, L.; Li, G.Y.; Larsson, E.G.; Caire, G. Is NOMA Efficient in Multi-Antenna Networks? A Critical Look at Next Generation Multiple Access Techniques. *arXiv* **2021**, arXiv:2101.04802.