Received August 13, 2017, accepted September 7, 2017, date of publication September 18, 2017, date of current version October 25, 2017.

Digital Object Identifier 10.1109/ACCESS.2017.2753380

Non-Orthogonal Multiple Access for Large-Scale 5G Networks: Interference Aware Design

KONPAL SHAUKAT ALI, (Student Member, IEEE), HESHAM ELSAWY, (Member, IEEE), ANAS CHAABAN, (Member, IEEE), AND MOHAMED-SLIM ALOUINI, (Fellow, IEEE)

Computer, Electrical, and Mathematical Sciences and Engineering Divison, King Abdullah University of Science and Technology, Thuwal 23955-6900, Saudi Arabia

Corresponding author: Konpal Shaukat Ali (konpal.ali@kaust.edu.sa)

ABSTRACT Non-orthogonal multiple access (NOMA) is promoted as a key component of 5G cellular networks. As the name implies, NOMA operation introduces intracell interference (i.e., interference arising within the cell) to the cellular operation. The intracell interference is managed by careful NOMA design (e.g., user clustering and resource allocation) along with successive interference cancellation. However, most of the proposed NOMA designs are agnostic to intercell interference (i.e., interference from outside the cell), which is a major performance limiting parameter in 5G networks. This paper sheds light on the drastic negative-impact of intercell interference on the NOMA performance and advocates interference-aware NOMA design that jointly accounts for both intracell and intercell interference. To this end, a case study for fair NOMA operation is presented and intercell interference mitigation techniques for NOMA networks are discussed. This paper also investigates the potential of integrating NOMA with two important 5G transmission schemes, namely, full duplex and device-to-device communication. This is important since the ambitious performance defined by the third generation partnership project for 5G is foreseen to be realized via seamless integration of several new technologies and transmission techniques.

INDEX TERMS 5G mobile communication, interference, NOMA.

I. INTRODUCTION

Non-orthogonal multiple access (NOMA) is promoted as one of the promising technologies to be adopted in 5G cellular networks. In contrast to conventional orthogonal multiple access (OMA), NOMA is foreseen to increase the network capacity by improving the spatial utilization of the scarce spectrum. Conventionally, temporal, spectral, and/or spatial orthogonalization are adopted to avoid intracell interference (i.e., interference among UEs in the same cell), which permits only one user-equipment (UE) to be served per time-frequency resource-block per BS. In NOMA cellular networks, multiple UEs can be clustered and simultaneously served over the same time-frequency resource-block. Particularly, NOMA UEs share the same time-frequency resources by either having their messages superposed in the power domain or in the code domain.[1] The mutual intracell interference among the clustered UEs is managed by proper resource allocation (RA) (e.g., transmission power and rate) in conjunction with successive interference cancellation (SIC) [1]. Consequently, UE clustering and RA are crucial for NOMA operation in order to minimize the effect of intracell interference [1], [2].

Motivated by its potential to increase the capacity of cellular networks, the design of NOMA transmission via UE clustering and RA has received much attention from the research community [3]–[7]. For instance, the importance of power allocation (PA) for achieving transmission rate gains in a single-cell NOMA downlink is emphasized in [4]. The effect of PA on transmission rate fairness among NOMA UEs is manifested in [5]. UE clustering is considered in [6] where selection strategies are proposed based on the distance between the UEs and the BS. In [7], it is shown that having NOMA UEs with more distinct channels enhances NOMA gains in a single-cell downlink scenario. The coexistence of a cognitive secondary NOMA system with a primary OMA network is studied in [8]. However, the study in [8] is limited to a single secondary NOMA BS. Tabassum *et al.* [9] study a large-scale uplink NOMA system where UEs are modeled using a cluster-process. A fixed number of UEs is assumed in a cluster and constant transmission power is considered.

In addition to the standalone NOMA networks studied in [3]–[7] and [9], the integration of NOMA with other 5G candidate technologies are also studied in the literature. Such studies are important to realize the 5G network, which

[1]There are other NOMA schemes such as pattern division multiple access (PDMA), bit division multiple access (BDMA), and interleave-division multiple access (IDMA) [1].

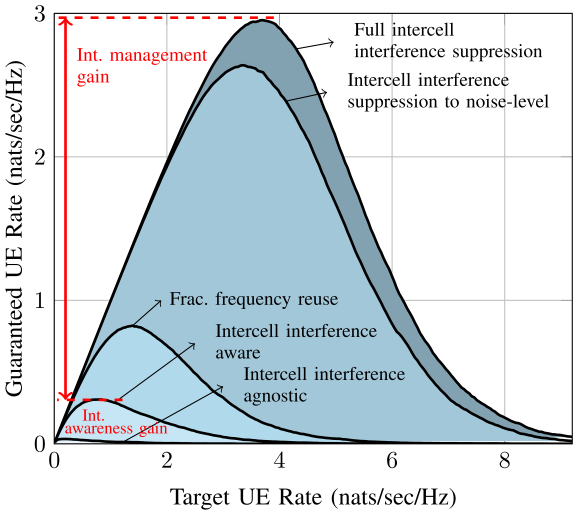


FIGURE 1. Downlink guaranteed-rates per UE for different levels of intercell interference in a NOMA setup with $N = 2$. Here, the guaranteed-rate is the minimum effective-rate of a UE in the setup, where the effective-rate of a UE is the target-rate times the probability that no outage occurs.

is foreseen to adopt several new transmission schemes. For instance, an integrated NOMA and FD transmission scheme is proposed in [10] to reduce relaying delay and improve the end to end throughput. Zhang *et al.* [11] exploit FD relaying to improve the reliability of NOMA transmissions, in which the NOMA UE with stronger channel simultaneously receives it own message and relays an older message designated to the UE with the weaker channel. A NOMA framework between FD BSs and half-duplex UE clusters is proposed in [12], which extends the well known FD 3-node topology to NOMA clusters [13]–[15]. Elbamby *et al.* [16] propose a selection criterion between NOMA and FD communication that is based on traffic conditions, network density, and self-interference cancellation capabilities. Applying NOMA to D2D communications is considered in [17]. The coexistence between NOMA-based cellular MU-MIMO and OMA D2D network is studied in [18]. However, the work in [10]–[12], [17], and [18] is limited to a single-cell scenario and the study in [16] considers a small multi-cell network.

Most of the aforementioned NOMA studies are myopic in the sense that each BS independently utilizes the local information (e.g., channel state information (CSI), UE locations, target-rates, power constraints, etc.) to perform UE clustering and RA. Such myopic schemes ignore intercell interference (i.e., interference from other cells), which is a fundamental performance limiting parameter in cellular networks. To illustrate the effect of intercell interference on NOMA rate, we show Fig. 1, which plots guaranteed versus target UE rates for a two-UE fixed-rate NOMA transmission in a large-scale downlink cellular network. The simulation environment considered to plot Fig. 1 is detailed in the Appendix. As shown in the figure, the guaranteed UE rate is always less than the target-rate due to transmission outages, which occur when the fading, noise, and interference lead to a channel capacity less than the fixed-transmission rate. Also, we observe the existence of an optimal target-rate that balances the tradeoff between outage probability and channel utilization. The figure shows that the guaranteed-rate becomes significantly less than the target-rate for intercell interference agnostic NOMA design. However, intercell interference awareness and management can significantly reduce outages and improve the UE guaranteed-rate, which highlights the drastic effect of intercell interference.

Having highlighted the importance of interference-aware NOMA design for 5G cellular networks, it remains to note that there are multiple challenges associated with such design such as:

- **Interference's stochastic nature:** This is due to several sources of uncertainty such as channel gains, network geometry, transmission powers, traffic requirements, etc., which make interference hard to estimate.
- **Location dependence:** Interference is highly dependent on the location of UEs within a BS's service area: UEs near the cell boundary experience more intercell interference.
- **Dense multi-tier topologies:** Cellular networks are evolving towards an ultra-dense multi-tier topology with irregular cell structure. Hence, the distance between a UE and its serving BS cannot be used to infer its location with respect to the cell boundary.
- **Dominant interferer:** Clustered NOMA UEs may have different dominant interfering sources (i.e., other transmitting UEs or BSs) due to their different spatial locations. This complicates the interference management process.

To the best of the authors' knowledge, such challenges are not addressed in the literature in the context of NOMA cellular networks.

Motivated by this, our article focuses on the design of *interference-aware NOMA schemes* in *large-scale 5G cellular networks*, where both intracell and intercell interference are jointly considered. The article first presents the different design objectives for NOMA cellular networks and discusses their advantages and disadvantages. Then, NOMA design for downlink and uplink scenarios are explicitly presented, in which the drastic impact of intercell interference on PA and UE sorting and clustering is discussed. In particular, we show that ignoring intercell interference devastates the performance, while interference suppression (which requires interference awareness) can result in tremendous gains. Since intercell interference has such a large impact on performance, we discuss intercell interference management techniques for NOMA cellular networks. Furthermore, as 5G will integrate several new technologies to achieve the envisioned gains in terms of rate, we highlight the potential of integrating NOMA with other 5G candidate technologies such as FD and D2D communication. We discuss the different kinds of interferences that arise due to each integration, their impact on the network, and possible techniques to handle them.

II. INTERFERENCE-AWARE DESIGN FOR NOMA NETWORKS

There are different types of performance criteria (utility) that can be used to assess the performance of a network. The different utilities impose a trade-off between performance, optimization complexity, and signaling overhead. There are also different optimization frameworks that impose a trade-off between the value of the overall network utility and the fairness among UEs, as shown Table 1. As such, the first step in NOMA design is to define the utility function and the optimization objective, based on the operator key performance indicator.

A. UTILITY FUNCTION

The utility function is usually related to the transmission rate, which is a function of the signal-to-interference-plus-noise-ratio (SINR). Consider N NOMA UEs per cell. The SINR associated with the i^{th} strongest UE is determined as

$$\text{SINR}_i = \frac{P_i \mathcal{F}_i}{\sigma_i^2 + \mathcal{I}_i + \mathcal{C}_i} \tag{1}$$

where P_i is the transmit power related to the i^{th} NOMA UE, $\mathcal{F}_i$ is the fading (distance dependent large-scale and multi-path small-scale) power gain, σ_i^2 is the noise power, $\mathcal{I}_i$ is the intercell interference power, and $\mathcal{C}_i$ is the intracell interference power.

If interference is treated as noise, Shannon's capacity specifies the maximum rate that can be reliably transmitted is

$$\begin{aligned} \mathcal{T}_i &= \log(1 + \text{SINR}_i) \\ &= \log(1 + \frac{P_i \mathcal{F}_i}{\sigma_i^2 + \mathcal{I}_i + \mathcal{C}_i}) \end{aligned} \tag{2}$$

Since the channel gains and interference may randomly change over time, the UE clustering, PA, and transmission rates should be continuously adapted to the instantaneous values of $\mathcal{F}_i$, $\mathcal{I}_i$, and $\mathcal{C}_i$ for all $i \in \{1, \ldots, N\}$. In this case, the utility is defined in terms of the ergodic rate $\mathcal{T}_i^{\text{erg}} = \mathbb{E}[\mathcal{T}_i]$, which necessitates high overhead to feedback the channel gains and the interference related information for each UE. Since such overhead is not affordable in large-scale networks, the utility should be defined according to some fixed transmission rates that are vulnerable to outage. A per-UE fixed transmission rate can be selected as $\log(1 + \theta_i)$, which is subject to decoding outages if $\text{SINR}_i < \theta_i$. Fixed rate transmissions reduce the required signaling overhead at the expense of an effective-rate of

$$\mathcal{T}_i^{\text{eff}} = \mathbb{P}\{\text{SINR}_i > \theta_i\} \log(1 + \theta_i), \tag{3}$$

which is less than the target-rate. Hence, UE clustering, PA, and target-rates should be carefully selected, according to interference and channel gain statistics, to maximize the effective-rates (cf. Fig. 1). These statistics can be obtained offline via stochastic geometry analysis [19]. A simpler approach is to define a global rate utility for all UEs (i.e., $\log(1 + \theta)$) and focus on the UE clustering and PA only, which simplifies the network design. The aforementioned utility definitions, which are summarized in Table 1, lead to different utility formulations that impose a tradeoff between performance and optimization complexity.

B. OPTIMIZATION OBJECTIVE

Having specified the utility function, the next step is to determine the optimization framework. In this paper we consider three types of optimization frameworks:

1) TOTAL UTILITY MAXIMIZATION

The total utility maximization represents an extreme scenario for the optimization formulation, which does not prioritize individual per-UE rate but only the total cell rate. Such formulation allocates all the resources to the UE with the best channel, which maximizes SINR, and hence, the effective-rate. In this extreme case, all UEs other that the UE with the strongest channel have zero rates and both NOMA and OMA would have the same total utility.

2) TOTAL UTILITY MAXIMIZATION WITH QUALITY OF SERVICE (QOS) CONSTRAINT

A QoS constraint is added to the total utility maximization formulation to guarantee a minimum effective-rate for all UEs. In this case, the optimization framework will operate all but the UE with the best channel at the QoS constraint and allocate all of the remaining resources (e.g., power) to the best UE.

3) UNIVERSAL FAIRNESS

Universal fairness among the UEs can be achieved via a max-min optimization framework which will achieve the maximum possible symmetric effective-rate for all UEs.

The different optimization frameworks are summarized in Table 1. Fig. 2 depicts a realization of each of these optimization frameworks in terms of the rate utility of each UE. As can be seen in the figure, the total utility maximization framework results in all resources being given to the strongest UE, i.e. UE1. Consequently it has the highest rate utility while all other UEs have zero rate utility. The total utility maximization with QoS constraint framework on the other hand ensures all UEs meet the QoS constraint and then dedicates the remaining resources to UE1 to maximize its utility. In the universal fairness framework all UEs are guaranteed identical (symmetric) effective-rate and the resources are distributed so that the maximum symmetric effective-rate is achieved. It ought to be noted that the total utility after applying the QoS constraints is lower than that of the unconstrained utility maximization which allocates all resources to UE1. Enforcing the symmetric effective-rate (i.e. universal fairness) further reduces the total utility than both of the aforementioned schemes. After specifying the design objective, several design problems must be tackled based on the selected objective.

TABLE 1. Summary of different utilities and objectives. We denote the vector of powers allocated to transmissions by P, and SINRs associated with target-rates by θ, with $\theta = (\theta_1, \ldots, \theta_N)$.

Scenario	UE Utility function	Max. Total Utility	Max. Total Utility s.t. QoS	Max. Fairness	Advantages	Disadvantages
Static Channel	$\mathcal{T}_i = \log(1 + \mathrm{SINR}_i)$	$\max_{\boldsymbol{P}} \sum \mathcal{T}_i$	$\max_{\boldsymbol{P}} \sum \mathcal{T}_i$ s.t. $\mathcal{T}_i \geq \mathrm{QoS}$	$\max_{\boldsymbol{P}} \min_i \mathcal{T}_i$	simple to study	impractical
Fading Channel Adaptive Rate	$\mathcal{T}_i^{\mathrm{erg}} = \mathbb{E}[\mathcal{T}_i]$	$\max_{\boldsymbol{P}} \sum \mathcal{T}_i^{\mathrm{erg}}$	$\max_{\boldsymbol{P}} \sum \mathcal{T}_i^{\mathrm{erg}}$ s.t. $\mathcal{T}_i^{\mathrm{erg}} \geq \mathrm{QoS}$	$\max_{\boldsymbol{P}} \min_i \mathcal{T}_i^{\mathrm{erg}}$	1) practical 2) max. utility	high overhead (instantaneous CSI)
Fading Channel Per-UE target-rate	$\mathcal{T}_i^{\mathrm{eff}} = \mathbb{P}\{\mathrm{SINR}_i > \theta_i\} \times \log(1 + \theta_i)$	$\max_{\boldsymbol{\theta}, \boldsymbol{P}} \sum \mathcal{T}_i^{\mathrm{eff}}$	$\max_{\boldsymbol{\theta}, \boldsymbol{P}} \sum \mathcal{T}_i^{\mathrm{eff}}$ s.t. $\mathcal{T}_i^{\mathrm{eff}} \geq \mathrm{QoS}$	$\max_{\boldsymbol{\theta}, \boldsymbol{P}} \min_i \mathcal{T}_i^{\mathrm{eff}}$	1) low overhead 2) requires CSI stats.	1) outage prone 2) low utility
Fading Channel Global target-rate	$\tilde{\mathcal{T}}_i^{\mathrm{eff}} = \mathbb{P}\{\mathrm{SINR}_i > \theta\} \times \log(1 + \theta)$	$\max_{\theta, \boldsymbol{P}} \sum \tilde{\mathcal{T}}_i^{\mathrm{eff}}$	$\max_{\theta, \boldsymbol{P}} \sum \tilde{\mathcal{T}}_i^{\mathrm{eff}}$ s.t. $\tilde{\mathcal{T}}_i^{\mathrm{eff}} \geq \mathrm{QoS}$	$\max_{\theta, \boldsymbol{P}} \min_i \tilde{\mathcal{T}}_i^{\mathrm{eff}}$	1) low overhead 2) requires CSI stats.	1) outage prone 2) low utility

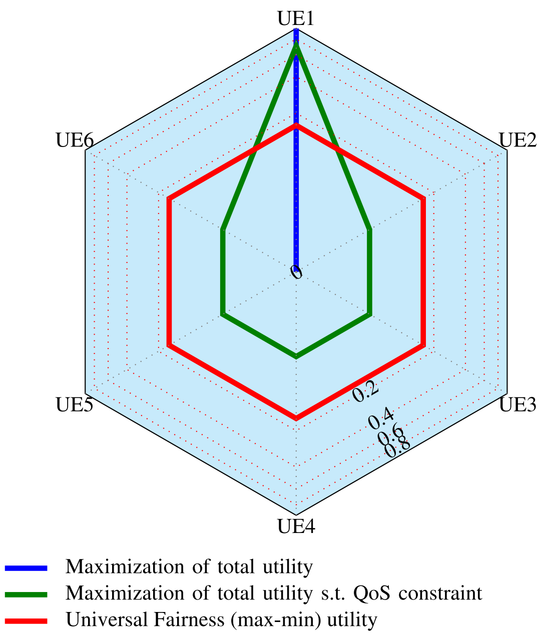


FIGURE 2. Rate utility of NOMA for different scenarios from Table 1. UEs are labeled from the strongest channel (i.e., UE1) to the weakest (i.e., UE6).

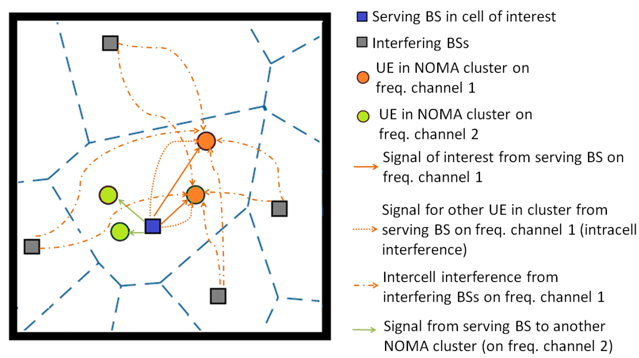


FIGURE 3. A cell of interest in a downlink NOMA setup with $N = 2$ and two UE clusters operating on different frequency channels. The intended signal and interferences experienced at the UEs of the cluster operating on frequency channel 1, and intended signals at the UEs of the cluster on frequency channel 2 are shown.

C. UE CLUSTERING

One design problem concerns forming the NOMA UE clusters and determining their sizes. The goal is that UEs within the same cluster operate in NOMA within the same time-frequency resource-block. Different clusters within the same BS are assigned orthogonal resources, and hence, intracell interference management is only required among UEs within the same cluster. However, clusters across different cells may share the same time-frequency resource block, which induces mutual intercell interference. This is illustrated in Fig. 3 which shows two different two-UE clusters operating on different frequency channels in the cell of interest. In particular, we show all the signals (intended and interfering) that the UEs on channel 1 receive and only the intended signal for the UEs on channel 2. Also shown are some nearby interfering BSs. As shown in the figure, the NOMA cluster in the cell of interest on frequency channel 1 does not receive interference from the signal of the serving BS to the cluster on channel 2. This cluster does, however, receive intercell interference from the signals of all other BSs on frequency channel 1. Also shown is the intracell interference from the signal intended for the other UEs in the NOMA cluster as the serving BS superposes the messages intended for all UEs in a cluster.

UE clustering is commonly based on sorting UEs in terms of their channel qualities [3]–[5] or link distance [6], [20] (discussed in the next subsection). However, clustering should not only incorporate the UE's local information (distance from BS, channel gain, target-rate, etc.), but also the intercell interference. Accounting for intercell interference can change the UEs' effective channel qualities, and consequently their sorting. Intercell interference also affects clustering from a different perspective. Namely, clustering UEs that share the same dominant intercell interferer facilitates further interference management, which again highlights the importance of interference-aware design.

D. UE SORTING

SIC is an important ingredient in NOMA for mitigating the intracell interference. The intracell interference, which arises within the cell, is comprised of the interference from messages that can not be canceled in the SIC chain as well as imperfections in SIC leading to residual successive interference (SI) from the canceled messages. The decoding order for SIC in uplink and downlink NOMA is not the same. In particular, a UE in the downlink decodes the messages of all weaker UEs (i.e. UEs after itself in the ordering) before decoding its own message. In the uplink on the other hand, a message is decoded after all messages from stronger UEs (i.e. UEs before itself in the ordering) have been decoded. Hence intracell interference can be written in the downlink as

$$\mathcal{C}_i = \mathcal{F}_i \left(\sum_{j=1}^{i-1} P_j + \beta \sum_{j=i+1}^{N} P_j \right) \tag{4}$$

and in the uplink as

$$\mathcal{C}_i = \sum_{j=i+1}^{N} P_j \mathcal{F}_j + \beta \sum_{j=1}^{i-1} P_j \mathcal{F}_j. \tag{5}$$

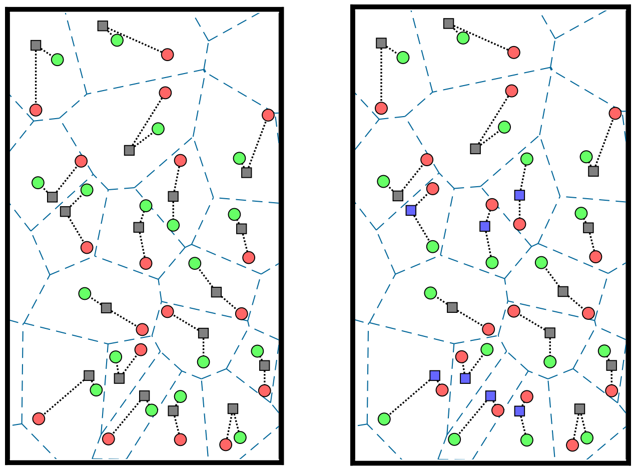
The factor β reflects SIC efficiency; perfect SIC implies that $\beta = 0$. As mentioned in the last subsection, UEs are often sorted in terms of either their channel gain or link distance. However, to obtain good results, UEs must be sorted in descending order of a better channel quality indicator. One possible such indicator is the ratio $\mathcal{G}_i = \frac{\mathcal{F}_i}{\mathcal{I}_i + \sigma_i^2}$, which accounts for the UE's channel taking into account the intercell interference.

Intercell interference $\mathcal{I}_i$ can drastically change the UEs sorting as shown in the example in Fig. 4 where a network with two-UE NOMA clusters is considered. The UEs with better channel quality indicator in each cell are marked in green and the weaker UEs in red. In Fig. 4a UEs are sorted according to channel gain and hence do not take into account intercell interference. However, in Fig. 4b the UEs are ordered according to intercell-interference based channel quality. We observe here that the cells with BSs marked in blue have the order of their UEs reversed from that in Fig. 4a. This is attributed to the irregular cell structure which can cause UEs with the best channel gain to the serving BS to have the highest interference-to-noise ratio. Thus, intercell interference has a detrimental effect in UE sorting and must be taken into account.

E. POWER/RATE ALLOCATION

Due to UE-ordering based SIC and corresponding intracell interference of the NOMA scheme (as can be seen from (4) and (5)), one should allocate transmission powers and rates such that: leftmargin=*

- In the downlink, UE i can decode and cancel messages designated to UEs $i+1, \ldots, N$ and tolerate interference from UEs $1, \ldots, i-1$.
- In the uplink, the BS can decode the messages with the allocated rates successively from the message of UE 1 to that of UE N, every time canceling the decoded message from its received signal.



(a) (b)

FIGURE 4. UE order in a two-UE NOMA cellular network, where squares, green circles, and red circles denote the BSs, stronger UEs, and weaker UEs. The Voronoi tessellation represents the coverage region of each BS and the dotted links represents association. The BSs affected by interference-aware sorting are highlighted in blue. (a) Interference-agnostic. (b) Interference-aware.

This results in allocation of higher power and/or lower rate for weaker UEs in the downlink and stronger UEs in the uplink.

As a consequence of the use of SIC in NOMA, the condition of successful decoding of a message is transformed into a joint condition wherein successful decoding of a message depends on successful decoding of the preceding messages in the SIC chain. In particular, the coverage expression for the i^{th} UE can be written as follows

$$\mathcal{P}_i = \mathbb{P}\left(\text{SINR}_{i,j} \geq \theta_j \forall j \in \mathcal{D}\right), \tag{6}$$

where $\text{SINR}_{i,j}$ is the SINR of the j^{th} message at the i^{th} decoder,[2] θ_j is the target-SINR corresponding to the target-rate of the j^{th} message, and $\mathcal{D} = \{1, \ldots, i\}$ for the uplink and $\mathcal{D} = \{i, \ldots, N\}$ for the downlink. It ought to be noted that (6) is different from the coverage used in (3) as SIC-based coverage results in a joint event. Additionally, the optimal power and rate allocation have to be done while taking intercell interference into account in order where RA done based on an intercell-interference agnostic network results in very poor network performance.

III. CASE STUDY

For the sake of exposition, we focus this section on the Max. Fairness utility (cf. Table 1) i.e. symmetric effective-rate for all UEs in the downlink and corresponding PA. In the uplink, we first show why employing PA that leads to symmetric effective-rates is challenging in a large network (detailed in Section III-B2). Full power transmissions, which do not lead to symmetric effective-rates, are then employed for the

[2] In the case of uplink there is only one receiver, namely the BS. This notation can be interpreted as the scenario where the BS is attempting to decode the i^{th} message and for this to be succesful all messages for UEs $1 < j < i$ also need to be decodable.

uplink. The impact of the interference on the rate utility and PA for the downlink and uplink scenarios are detailed. The operation scenario used in this case study and the rest of the article is detailed in the appendix. It ought to be highlighted that the operation scenario (in the appendix) uses a global target-rate.

A. IMPACT OF INTERCELL INTERFERENCE ON RATE UTILITY

Due to the SIC, successful decoding of a message at the i^{th} UE in the downlink is a joint event of decoding (for cancellations purpose) all messages for weaker UEs in addition to its own message. In the uplink, the BS is interested in all messages, and hence, successful decoding incorporates all messages starting with the strongest UE's message. Consequently, the effective-rate associated with the i^{th} NOMA UE is

$$\tilde{\mathcal{T}}_i^{\text{eff}} = \log(1+\theta)\mathcal{P}_i \tag{7}$$

It should be mentioned that since the optimization framework considered in the appendix is based on global target-rate, we have θ in (7), otherwise we would employ θ_i for calculating $\mathcal{T}_i^{\text{eff}}$. Similarly, due to the global target-rate, $\theta_j = \theta, \forall j$ in (6). The intercell interference is usually overlooked in the literature, as only small-scale (i.e., single-cell) NOMA scenarios are considered [3]–[5]. However, since NOMA is foreseen to be deployed in 5G networks, which are intrinsically ultra-dense and interference limited, neglecting intercell interference is not a justifiable practice as highlighted in Fig. 1.

As specified before, Fig. 1 is a plot of the guaranteed-rate (i.e. minimum effective-rate according to (7) for all UEs) against the target-rate in the downlink with $N = 2$. It ought to be mentioned that since symmetric effective-rates is the objective of our case study, the effective-rate is identical for all UEs except when intercell interference is not accounted for in PA (detailed in the next subsection). The figure depicts the achievable gains in guaranteed-rate for a UE as intercell interference mitigation is improved. The upper-limit for this is the plot of full interference supression, which, although impractical in reality, corresponds to no frequency reuse among BSs. Equivalently, it can be viewed as a single-cell setup. Fig. 1 thereby also emphasizes how studies on small-scale setups significantly overestimate actual guaranteed-rate and consequently network performance, stressing the importance of studying NOMA in the context of a real large-scale network.

B. IMPACT OF INTERCELL INTERFERENCE ON PA AND FAIRNESS

Since signals are multiplexed in the power domain, NOMA requires PA for each message given a power constraint.

1) DOWNLINK

In the downlink, a particular BS needs to transmit N messages using power P. Interestingly, when all messages are sent at the same target-rate, unlike conventional PA schemes such as water filling strategies, downlink NOMA-UEs with poor channel conditions are allocated larger power than those with stronger channels [2]. This is done to ensure that a UE can treat the messages intended for UEs with stronger channels than itself as noise. Assuming the same target-rate for all messages, to achieve symmetric effective-rates we require a PA scheme that equalizes SINR for each message at its respective receiver. This will ensure that coverage will be identical and due to a global target-rate, the effective-rate will be symmetric.

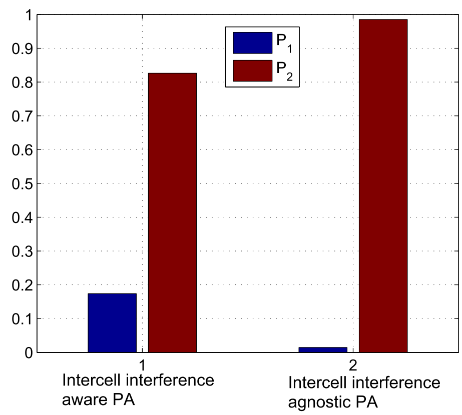


FIGURE 5. PA for UEs in a downlink NOMA setup with $N = 2$ in the intercell interference aware and agnostic cases.

Having a PA scheme that aims to achieve symmetric effective-rate without taking into account intercell interference in a large-scale network is far from optimal. In fact, a symmetric effective-rate is never achieved in this case because excluding intercell interference from the PA makes the SINR of stronger channels appear much larger than that of weaker channels. This results in a PA scheme that gives a lot more power to the weaker channels than it ought to. This is demonstrated in Fig. 5 where a two-UE downlink NOMA setup is considered. Intercell interference aware and agnostic PA are compared. We observe that intercell interference agnostic PA results in a 92% decrease in the allocated power of the stronger UE compared to its intercell interference aware counterpart, while the weaker UE has a 19% increase. In a large-scale downlink setup, the intercell interference agnostic approach deteriorates the performance of the UEs with stronger channels significantly as the impact of intercell interference dominates the impact of intracell interference and noise. Additionally, in a real network, such an intercell interference agnostic PA scheme is not able to provide fairness, i.e. a symmetric effective-rate, either. This in turn deteriorates the total effective cell rate (TECR), which is the sum of the individual effective rates of all the UEs in the NOMA cluster of a cell. We use the TECR as a measure of network performance. Incorporating intercell interference in the PA results in a less dramatic difference between powers allocated to the stronger and weaker channels, and better network performance.

In Fig. 1, we observe that the intercell interference agnostic PA guarantees a much lower effective-rate than its interference aware counterpart. This is due to the stronger UE getting significantly lower power than its requirement causing it to become the bottleneck of the guaranteed effective-rate in the agnostic case. This highlights that the large impact of intercell interference on the SINRs and therefore correct PA. This intercell interference can be taken into account by large-system analysis using stochastic geometry.

2) UPLINK

Although a more involved PA strategy can be used for uplink NOMA, it is difficult to equalize SINRs associated with each message (to achieve symmetric effective-rate) at the BS in this case. This is because intercell interference in the uplink comes from transmitting UEs from interfering cells, which like the UEs in the cell of interest require a PA strategy of their own. Game-theoretic approaches to find solutions to this may be employed. We do not delve into such approaches in this article and propose some simpler alternatives. One solution would be to overestimate the intercell interference by assuming that the interfering UEs transmit using their full power and then do PA that attempts to equalize SINR based on the overestimated intercell interference. Another approach is to enforce all uplink NOMA-UEs to transmit with their maximum power as done in [9].

PA based on overestimated intercell interference results in much worse network performance than full-power transmissions as shown in Fig. 6 where a two-UE uplink NOMA setup is considered. Dashed lines represent full-power transmissions and solid lines are for PA based on overestimated intercell interference. We observe that attempting to equalize SINR using overestimated intercell interference does not result in fairness i.e. in symmetric effective-rates for the UEs. In fact, the weaker UE performs only a little better than the full-power transmission case, while the stronger UE does much worse. This deteriorates the overall performance without achieving symmetric effective-rates in the case of overestimated intercell interference. It can be explained by the fact that the overestimated intercell interference, although impacts both UEs, makes the PA tilt in favor of the weaker UE which is not able to improve performance as much even with the additional resource. This too highlights the significance of intercell interference and how its inaccurate estimation deteriorates network performance. Additionally, it emphasizes that accounting for intercell interference inaccurately is worse than inefficient RA.

IV. INTERCELL INTERFERENCE MANAGEMENT

As highlighted in the previous sections and shown in Fig. 1, there is significant room for improving the NOMA performance via intercell interference management. Roughly, interference management techniques can be divided into two main categories: offline and online. Both techniques are briefly discussed in the sequel.

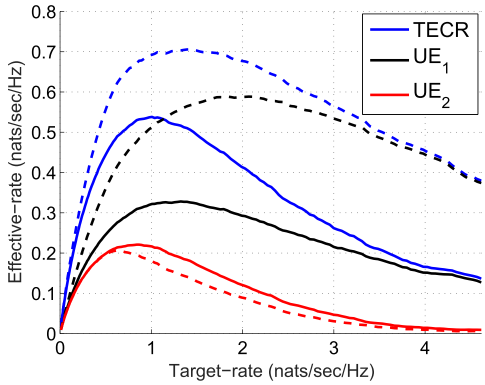


FIGURE 6. Total and individual UE effective-rates vs. target-rate for uplink-NOMA with $N = 2$. Dashed lines represent full-power transmission, solid lines represent PA based on attempted fairness when intercell interference is overestimated.

A. OFFLINE INTERFERENCE MANAGEMENT

The main advantage of the offline interference management techniques is that they require no operational overhead. However, the price paid is a rigid design that may lead to spectrum underutilization. Common examples of offline interference management techniques are

- **Frequency reuse:** the entire spectrum is partitioned and allocated to BSs such that neighboring BSs do not share the same set of channels [21]. Despite its fundamental role in previous generations of cellular networks, frequency reuse cannot be employed in 5G networks for the following reasons; i) it underutilizes the spectrum in less loaded cells; and ii) there is no efficient frequency reuse scheme that avoids interference among neighboring BSs due to the irregular structure of cells, especially due to dense multi-tier topologies.
- **Fractional frequency reuse (FFR):** cells are partitioned into cell-center and cell-edge regions. The spectrum is divided into two main chunks, the cell-center chunk and the cell-edge chunk. The cell-center chunk is universally reused over all cell-center regions. The cell-edge chunk is allocated to cell-edge regions such that neighboring cells do not share common cell-edge channels [21]–[24]. FFR leads to a better frequency utilization than conventional frequency reuse. However, the notion of cell-edge and cell-center should be based on the interference to noise ratio (INR), which depends on intercell interference, rather than the cell geometry.
- **Sectoring:** the service-area of each BS is divided into several sectors using directional antennas [25]. FFR among the sectors can be applied such that neighboring BSs' sectors do not share the same set of cell-edge channels. Due to irregular cell structure, it is hard to equalize the region covered by each sector while ensuring no cell-edge frequencies overlap among adjacent cells.

NOMA SPECIFIC CHALLENGES

The offline interference management techniques assign frequencies to geographical areas which restricts the NOMA clustering process and deteriorates the multi-user diversity. Furthermore, there could be regions over (under) populated with UEs more (less) than the affordable cluster size, which may lead to UE blocking (resource underutilization).

B. ONLINE INTERFERENCE MANAGEMENT

Online interference management is conducted by means of BSs cooperation and/or coordination, which may involve high signaling overhead and increases the RA complexity. Some examples of online interference management techniques are:

- **Silencing:** this is a simple form of online interference management, where the serving BS sends a silencing request, on a designated frequency band, to dominant interfering BSs and/or UEs [26]. However, this leads to spectrum underutilization in the cells of the silenced interferers.
- **Cognitive spectrum access:** BSs are divided into spectrum owners and cognitive BSs. Cognitive BSs can reuse the spectrum via underlay, overlay, or interweave techniques [27]. In the underlay scheme, a cognitive BS should know the CSI at the primary receiver and operate subject to an interference constraint. In the overlay, the cognitive BS allocates some of its power to aid the primary transmission and the rest of the power to its own transmission. The interweave technique is an opportunistic spectrum access technique where the cognitive BS uses the idle channels of the primary BSs.
- **Cooperative beamforming:** BSs employ multiple antennas and are divided into clusters. BSs with the same cluster perform joint precoding and PA to align the mutual interference at the scheduled UEs [22], [28]–[30]. Such a technique necessitates high signaling overhead for CSI feedback and imposes high RA complexity.
- **Cooperative multi-point transmission (CoMP):** UEs at the cell boundary between multiple BSs can be jointly served by these BSs in a network-MIMO fashion [24]. Such a technique significantly enhances the cell-edge UEs, however, at the expense of i) high signaling overhead for CSI; and ii) high backhaul utilization for sharing the UEs messages among the cooperating BSs.

NOMA SPECIFIC CHALLENGES

The main challenge imposed by NOMA for online interference management is that the NOMA UEs may have different dominant intercell interferers, which complicates the intercell interference management. Clustering the NOMA UEs that share common dominant intercell interference source may not be efficient from the intracell interference perspective. Balancing the tradeoff between intracell and intercell interference in NOMA clustering is a fundamental open problem. Determining which interferer to silence, for instance, can be challenging as decisions such as whether to silence the interferer that most deteriorates one receiver or silencing the interferer that is most detrimental to all receivers in a cluster in the downlink needs to be made. Similarly, in the uplink one needs to determine how many interferers from a cluster to silence and if this number can vary in different clusters. In other words, one needs to know which NOMA receiver/transmitter needs to be prioritized most. In the case of beamforming and CoMP interference management, the complexity of the precoding and RA, along with the CSI signaling overhead, increases with the cluster size. In the cognitive spectrum access scenario, there are multiple primary UEs that should be simultaneously considered by cognitive NOMA UEs. Consequently, the spectrum sensing complexity increases with NOMA cluster size.

V. INTEGRATION OF NOMA WITH OTHER 5G TECHNOLOGIES

5G defines a performance leap of 1000 times capacity when compared to current 4G networks. Such ambitious performance gain is foreseen to be fulfilled via integrating several key technologies [31], [32] including NOMA, FD, and D2D communication. This section highlights the opportunities and challenges for the integration of NOMA with the aforementioned technologies. Before delving into these details we specify the differences between our integration and what exists in the literature.

Often works that integrate the aforementioned technologies with NOMA simply employ them on different devices or at different times. They do not necessarily integrate the technologies in the sense of being deployed simultaneously by the same devices. In particular, in [16] either NOMA or FD are employed in a time frame and the two can not be employed simultaneously. In [10] and [11] the BS transmits using NOMA and one of the receivers relays an older message in a FD but non-NOMA fashion. Similarly, in [18] D2D and NOMA are not integrated by the same devices since the cellular MU-MIMO employs NOMA while overlaid with simple D2D communication.

We are interested in studying the impact of deploying these technologies with NOMA simultaneously at the same devices. Studies that integrate the technologies in such a way include [12]. Here the BS deploys NOMA and FD simultaneously by communicating with uplink and downlink NOMA clusters. The uplink and downlink clusters, however, do not operate in FD. It should be noted that [12] focuses on a single-cell scenario thereby not studying the impact of integrating the two technologies in a large network. Similarly, NOMA is employed by D2D devices in [17], thereby integrating the two technologies at the same devices. Again, however, the study is on a single-cell setup, thereby not accounting for the impact of the integration in a large network. This large network aspect is what we discuss next.

A. FULL-DUPLEX NOMA

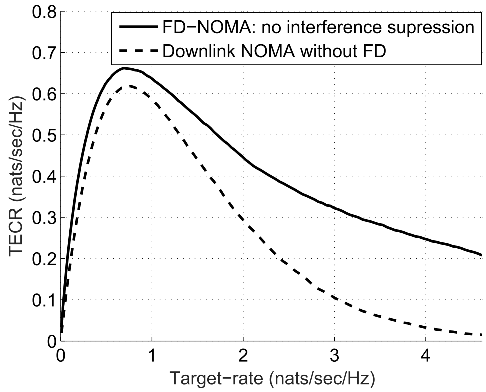
FD communication allows a transmit-receive pair to communicate simultaneously on the same channel, i.e. in the same time-frequency resource block [33], [34]. Conventionally, FD communication was rendered infeasible due to the overwhelming self-interference. Thanks to the recent advances in digital and analog circuit design, sufficient self-interference cancellation to operate in FD mode is now viable [35], [36].

Integrating NOMA with FD communication enables simultaneous uplink and downlink communication between the BS and all UEs within the NOMA cluster, which further improves the spectral utilization when compared to the standalone NOMA or FD scenarios. However, FD-NOMA communication increases the aggregate interference dramatically. Since the uplink and downlink operate on the same frequency band, FD-NOMA experiences the following sources of interference

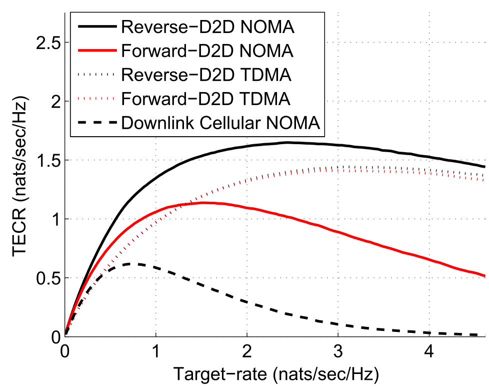
- **Intra-mode intercell interference:** downlink-to-downlink and uplink-to-uplink interference from other cells.
- **Inter-mode intercell interference:** downlink-to-uplink and uplink-to-downlink interference from other cells.
- **Intra-mode intracell interference:** downlink-to-downlink and uplink-to-uplink NOMA interference from transmitters within the same cluster in the same the cell.
- **Inter-mode intracell interference:** uplink-to-downlink only, which arises from other uplink transmissions of UES within the same cluster.
- **Residual self-interference:** due to the imperfection of self-interference cancellation.

Fig. 7a plots TECR against target-rate for a two-UE NOMA setup. Traditional downlink NOMA (without FD) is compared against FD-NOMA. The figure highlights the potential of harvesting rate gains via FD-NOMA without any interference management. However, the above discussion highlights that the integration between FD and NOMA requires sophisticated interference management techniques to alleviate the aforementioned types of interference. Note that alleviating one source of interference may aggravate another. For instance, inter-mode intracell interference management may be facilitated by clustering UEs that are sufficiently separated across the cell. However, sufficiently separated UEs may have different dominant intercell interferers, which complicates the intercell interference management.

Fig. 8 is a plot of TECR against target-rate for the two-UE FD-NOMA setup with different interference suppression. We observe from the figure that suppressing intercell interference has the most significant impact on FD-NOMA performance as only a 30% intercell interference suppression results in significant gains in TECR. Techniques to mitigate intercell interference may require NOMA UEs to be clustered in order to share common strong interferers for instance. However, at the UEs, the inter-mode intracell interference impacts the SINR experienced significantly more than any other interference source. As observed in Fig. 8, suppressing intra-mode intracell interference (without any intercell interference suppression), which only impacts the downlink, also results in approximately the same TECR improvement as 30% intercell interference suppression. This can be explained by the fact that the source of the inter-mode intracell interference, i.e. the UE cluster, occurs inside the cell thereby having the most detrimental impact on downlink performance. Suppressing this interference, by imposing a certain minimum distance between the UEs in a cluster for instance, is therefore of utmost importance for downlink performance. This sheds light on the need to prioritize dealing with interference sources; suitable intercell interference



(a)



(b)

FIGURE 7. Potentials of integrating NOMA with 5G technologies for $N = 2$ case. (a) TECR vs. target-rate for FD-NOMA. (b) TECR vs. target-rate for D2D-NOMA.

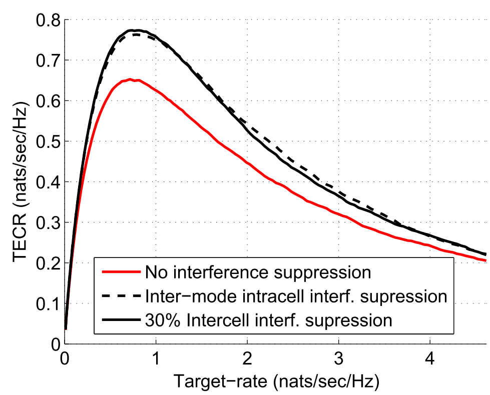


FIGURE 8. TECR vs. target-rate for FD-NOMA with different interference suppression when $N = 2$.

mitigation techniques can then be employed to improve the network performance further.

Despite the imposed interference challenges, we observe the potential of FD-NOMA to improve the rate from Figs. 7a and 8. Additionally, from Fig. 8 we can extrapolate that intercell interference is overall most detrimental to FD-NOMA network performance.

B. D2D NOMA

D2D communication allows proximate UEs to bypass BSs and communicate directly in a peer-to-peer fashion. Such short range direct D2D communication is foreseen to relieve BS congestion, improve spatial spectrum utilization, reduce power consumption, and reduce latency [37], [38]. Integrating NOMA with D2D enables one-to-many (denoted as forward-D2D) and many-to-one (denoted as reverse-D2D) communication, which can further improve the D2D gains. To be specific, forward-D2D involves NOMA transmission from one D2D transmitter to a cluster of D2D receivers, while reverse-D2D involves NOMA transmission from a cluster of D2D transmitters to one D2D receiver. Note that D2D-UEs operate under a limited power budget when compared to BSs, and hence, the induced interference from D2D-NOMA communication can be affordable.

The D2D-NOMA communication can either share the uplink or downlink resources of the cellular communicants. For the sake of exposition, we consider a D2D-NOMA framework sharing a downlink OMA cellular transmission. Additionally, we assume only one D2D-NOMA cluster in each cell on the resource-block being considered.

1) FORWARD-D2D NOMA

The forward-D2D NOMA experiences the following interference sources:

- **Cellular interference** from all cellular interferers, which can be divided into intracell and intercell according to the D2D cluster position.
- **Inter-cluster D2D interference** from transmitters of other D2D-NOMA clusters.
- **NOMA interference** from the messages designated to other D2D UEs within the same cluster.

2) REVERSE-D2D NOMA

The reverse-D2D NOMA experiences the following interference sources:

- **Cellular interference** which can be divided into intracell and intercell according to the D2D cluster position.
- **Inter-cluster D2D interference** from transmitters of other D2D-NOMA clusters.
- **NOMA interference** from other D2D UEs within the same cluster.

Fig. 7b is a plot of TECR against target-rate. We plot both forward and reverse D2D-NOMA with cluster size $N = 2$. These are compared against forward and reverse D2D-TDMA (OMA) where the two-UE cluster shares time resources. For completeness, we also plot traditional downlink cellular NOMA with cluster size $N = 2$. Despite the increased interference, Fig. 7b depicts the potential of harvesting rate gains from D2D-NOMA. We observe that the reverse D2D-NOMA always outperforms D2D-TDMA. However, forward D2D-NOMA outperforms the D2D-TDMA prior to the 5 dB, after which the case is reversed. This is because in the forward D2D-NOMA the transmitting D2D-UE splits its limited power budget among its receivers. In contrast, the reverse D2D-NOMA UEs use their full power budget for transmission. Hence, it is recommended to use TDMA in the forward D2D link if high target-rate is required.

Fig. 9 plots TECR against target-rate for the two-UE forward-D2D NOMA setup with different interference suppression. Suppressing cellular intracell interference (i.e. from the BS) at the D2D receivers is of utmost importance as it has the most significant impact on their performance. This is due to the high transmit power of BSs and small distance from the serving BS inside the cell. From the figure we observe that suppressing cellular intracell interference offers significant gains from the case without any interference suppression. Additionally, the severity of this interference is reflected in the fact that suppressing it leads to similar gains as an intercell interference suppression of 65% (compared to only a 30% intercell interference suppression in the FD-NOMA case), emphasizing the importance of managing it. Of course intercell interference still has the largest impact on TECR as its suppression leads to the largest gains. Suppressing cellular intracell interference may require techniques such as scheduling D2D receivers at a certain distance from the BS in the cell. Unlike the case of FD-NOMA, this complies with techniques to enhance intercell interference management by clustering the D2D UEs far from the BS for instance. Since intercell interference impacts the performance significantly as well, intercell interference management techniques should be employed in addition to cellular intracell interference management techniques.

Similarly for the case of reverse-D2D NOMA, the cellular intracell interference (i.e. from the BS inside the cell) to the D2D receiver has the most significant impact as the BS transmits at high power and is inside the cell. Additionally, although low power, the multiple transmitting D2D UEs cause significant interference at the cellular receiver as they may be large in number depending on the cluster size and lie inside the cell. Like its forward-D2D counterpart, interference between the intracell transmitters and receivers needs to be handled in addition to intercell interference. The results for this are similar to the forward-D2D NOMA case and are not included for the sake of exposition.

VI. SUMMARY

In this article we demonstrate the negative impact of intercell interference on non-orthogonal multiple access (NOMA) performance in 5G networks. We also present an interference aware NOMA design that jointly accounts for intercell and intracell interference. Particularly, we discuss interference

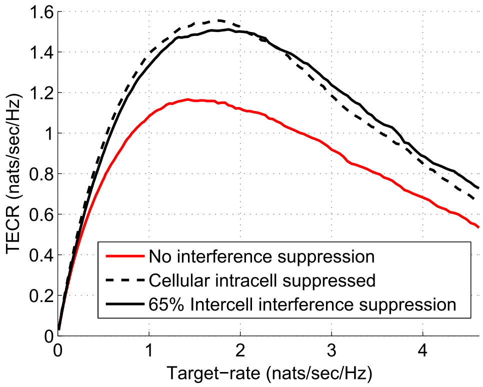


FIGURE 9. TECR vs. target-rate for forward-D2D NOMA with $N = 2$ and different interference suppression.

aware UE sorting along with the different interference aware design objectives and optimization frameworks. To this end, a case study for symmetric NOMA transmission rate utility for universal fairness in the downlink and full-power transmission in the uplink, is presented and interference mitigation techniques in the context of NOMA are highlighted. Last but not least, we consider the potentials and challenges of integrating NOMA with other 5G candidate technologies, namely, full-duplex and device-to-device communication, in a large network. The different kinds of interferences that arise, their impact, and how they can be handled are discussed. The main recommendations we make in this article can be summarized as follows:

- Intercell interference must be accounted for to avoid significantly overestimating network performance (single-cell).
- Downlink NOMA PA should account for intercell interference to avoid drastically deteriorating performance.
- Uplink NOMA performs better using full-power transmissions rather than PA based on overestimated intercell interference. This highlights that accounting for intercell interference inaccurately damages performance significantly more than suboptimal RA.
- Intercell interference awareness can change the UE sorting order to a more efficient one.
- Accounting for intercell interference can improve UE clustering which plays a role in improving the efficiency of interference management techniques.
- Intercell interference by far has the most detrimental impact on network performance. However, when NOMA is integrated with other technologies, certain receivers may be more severely impacted by another interference source and need to be handled accordingly.
  - In FD-NOMA the UEs are drastically impacted by inter-mode intracell interference. Techniques to manage this may contradict managing intercell interference. However, handling this should be prioritized first followed by other methods to handle intercell interference.
  - In D2D-NOMA the D2D receiver(s) are most impacted by the cellular intracell interference. Techniques to manage this also help mitigate intercell interference and should be used alongside other intercell interference management techniques.

All the above points represent interesting problems for future investigation, which can lead to practical design guidelines for NOMA systems.

APPENDIX
OPERATION SCENARIO

The operation scenario used in this work is detailed below.

- A homogeneous Poisson Point Process (PPP) cellular network of intensity $\lambda = 10$ BSs/km^2 is employed. The PPP assumption for the locations of BS have several practical and theoretical validations [19].
- N UEs are dropped uniformly in the Voronoi cell of each BS (cf. Fig. 4).
- A Rayleigh fading environment and a power law path-loss model, where the signal decays with distance r as $r^{-\eta}$ and the path-loss exponent $\eta = 4$, are assumed.
- We set the power budgets to $P = 1$ W and $P_u = 0.2$ W for the BS and UEs, respectively.
- In the case of NOMA, the entire time-frequency block is used for transmission. In the case of OMA, the time resource (i.e. time-division multiple access (TDMA)) is split between the N UEs.
- Fixed rate transmission is employed and a global target-rate (i.e., the same target-rate for all UEs in a cluster) is used.
- Fixed rate transmission is vulnerable to outage, correspondingly the effective-rate is defined as coverage probability multiplied by target-rate. For NOMA, this is calculated using (7), which employs the coverage probability given in (6). Effective-rate for OMA is calculated using the general formula for effective-rate in (3) multiplied by the fraction of the resource being shared (time fraction for TDMA).
- PA and optimization objective for NOMA:
  - Downlink: Optimization objective is universal fairness (i.e. symmetric effective-rate for all UEs). PA is done accordingly so that effective-rates can be identical. Since global target-rate is used, identical effective-rates are achieved by a PA that can equalize SINRs for the UEs.
  - Uplink: In the uplink, superposition is not required by a transmitter (UE). All UEs transmit with full-power and therefore do not achieve symmetric effective-rate. The superiority of this approach and the challenges associated with employing a PA that achieves symmetric effective-rate in the uplink are highlighted in Section III-B2.

Acknowledgment

The authors would like to thank Prof. Md. Jahangir Hossain for his insightful feedback on the manuscript.

REFERENCES

[1] 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.

[2] 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.

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

[4] Z. Ding, Z. Yang, P. Fan, and H. V. Poor, "On the performance of non-orthogonal multiple access in 5G systems with randomly deployed users," *IEEE Signal Process. Lett.*, vol. 21, no. 12, pp. 1501–1505, Dec. 2014.

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

[6] Y. Liu, Z. Ding, M. Elkashlan, and H. V. Poor, "Cooperative non-orthogonal multiple access with simultaneous wireless information and power transfer," *IEEE J. Sel. Areas Commun.*, vol. 34, no. 4, pp. 938–953, Apr. 2016.

[7] Z. Ding, P. Fan, and H. V. Poor, "Impact of user pairing on 5G nonorthogonal multiple-access downlink transmissions," *IEEE Trans. Veh. Technol.*, vol. 65, no. 8, pp. 6010–6023, Aug. 2016.

[8] Y. Liu, Z. Ding, M. Elkashlan, and J. Yuan, "Nonorthogonal multiple access in large-scale underlay cognitive radio networks," *IEEE Trans. Veh. Technol.*, vol. 65, no. 12, pp. 10152–10157, Dec. 2016.

[9] H. Tabassum, E. Hossain, and M. J. Hossain, "Modeling and analysis of uplink non-orthogonal multiple access in large-scale cellular networks using poisson cluster processes," *IEEE Trans. Commun.*, vol. 65, no. 8, pp. 3555–3570, Aug. 2017.

[10] C. Zhong and Z. Zhang, "Non-orthogonal multiple access with cooperative full-duplex relaying," *IEEE Commun. Lett.*, vol. 20, no. 12, pp. 2478–2481, Dec. 2016.

[11] Z. Zhang, Z. Ma, M. Xiao, Z. Ding, and P. Fan, "Full-duplex device-to-device-aided cooperative nonorthogonal multiple access," *IEEE Trans. Veh. Technol.*, vol. 66, no. 5, pp. 4467–4471, May 2017.

[12] Y. Sun, D. W. K. Ng, Z. Ding, and R. Schober, "Optimal joint power and subcarrier allocation for full-duplex multicarrier non-orthogonal multiple access systems," *IEEE Trans. Commun.*, vol. 65, no. 3, pp. 1077–1091, Mar. 2017.

[13] K. M. Thilina, H. Tabassum, E. Hossain, and D. I. Kim, "Medium access control design for full duplex wireless systems: Challenges and approaches," *IEEE Commun. Mag.*, vol. 53, no. 5, pp. 112–120, May 2015.

[14] A. AlAmmouri, H. ElSawy, and M.-S. Alouini, "Harvesting full-duplex rate gains in cellular networks with half-duplex user terminals," in *Proc. IEEE Int. Conf. Commun. (ICC)*, May 2016, pp. 1–7.

[15] K. Sundaresan, M. Khojastepour, E. Chai, and S. Rangarajan, "Full-duplex without strings: Enabling full-duplex with half-duplex clients," in *Proc. 20th Annu. Int. Conf. Mobile Comput. Netw. (MobiCom)*, 2014, pp. 55–66.

[16] M. S. Elbamby, M. Bennis, W. Saad, M. Debbah, and M. Latva-Aho, "Resource optimization and power allocation in in-band full duplex (IBFD)-enabled non-orthogonal multiple access networks," *IEEE J. Sel. Areas Commun.*, to be published, doi: 10.1109/JSAC.2017.2726218.

[17] J. Zhao, Y. Liu, K. K. Chai, Y. Chen, M. Elkashlan, and J. Alonso-Zarate, "NOMA-based D2D communications: Towards 5G," in *Proc. IEEE Global Commun. Conf. (GLOBECOM)*, Dec. 2016, pp. 1–6.

[18] H. Sun, Y. Xu, and R. Q. Hu, "A NOMA and MU-MIMO supported cellular network with underlaid D2D communications," in *Proc. IEEE 83rd Veh. Technol. Conf. (VTC)*, May 2016, pp. 1–5.

[19] H. ElSawy, A. Sultan-Salem, M.-S. Alouini, and M. Z. Win, "Modeling and analysis of cellular networks using stochastic geometry: A tutorial," *IEEE Commun. Surveys Tuts.*, vol. 19, no. 1, pp. 167–203, 1st Quart., 2016.

[20] 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.

[21] O. G. Aliu, A. Imran, M. A. Imran, and B. Evans, "A survey of self organisation in future cellular networks," *IEEE Commun. Surveys Tuts.*, vol. 15, no. 1, pp. 336–361, 1st Quart., 2013.

[22] N. Saquib, E. Hossain, L. B. Le, and D. I. Kim, "Interference management in OFDMA femtocell networks: Issues and approaches," *IEEE Wireless Commun.*, vol. 19, no. 3, pp. 86–95, Jun. 2012.

[23] N. Saquib, E. Hossain, and D. I. Kim, "Fractional frequency reuse for interference management in LTE-advanced hetnets," *IEEE Wireless Commun.*, vol. 20, no. 2, pp. 113–122, Apr. 2013.

[24] K. Yang, "Interference management in LTE wireless networks [Industry Perspectives]," *IEEE Wireless Commun.*, vol. 19, no. 3, pp. 8–9, Jun. 2012.

[25] D. Tse and P. Viswanath, *Fundamentals of Wireless Communications*. Cambridge, U.K.: Cambridge Univ. Press, 2004.

[26] G. Nigam, P. Minero, and M. Haenggi, "Spatiotemporal cooperation in heterogeneous cellular networks," *IEEE J. Sel. Areas Commun.*, vol. 33, no. 6, pp. 1253–1265, Jun. 2015.

[27] Q. Zhao and B. M. Sadler, "A survey of dynamic spectrum access," *IEEE Signal Process. Mag.*, vol. 24, no. 3, pp. 79–89, May 2007.

[28] L. Dong, A. P. Petropulu, and H. V. Poor, "Cooperative beamforming for wireless ad hoc networks," in *Proc. IEEE Global Commun. Conf. (GLOBECOM)*, Nov. 2007, pp. 2957–2961.

[29] M. H. Hassan and M. J. Hossain, "Cooperative beamforming for cognitive radio systems with asynchronous interference to primary user," *IEEE Trans. Wireless Commun.*, vol. 12, no. 11, pp. 5468–5479, Nov. 2013.

[30] T. Yi, L. Guo, K. Niu, H. Cai, J. Lin, and W. Ai, "Cooperative beamforming in cognitive radio network with hybrid relay," in *Proc. 19th Int. Conf. Telecommun. (ICT)*, Apr. 2012, pp. 1–5.

[31] K. S. Ali, H. ElSawy, and M.-S. Alouini, "Modeling cellular networks with full-duplex D2D communication: A stochastic geometry approach," *IEEE Trans. Commun.*, vol. 64, no. 10, pp. 4409–4424, Oct. 2016.

[32] J. Qiao, X. Shen, J. Mark, Q. Shen, Y. He, and L. Lei, "Enabling device-to-device communications in millimeter-wave 5G cellular networks," *IEEE Commun. Mag.*, vol. 53, no. 1, pp. 209–215, Jan. 2015.

[33] G. Liu, F. R. Yu, H. Ji, V. C. M. Leung, and X. Li, "In-band full-duplex relaying: A survey, research issues and challenges," *IEEE Commun. Surveys Tuts.*, vol. 17, no. 2, pp. 500–524, 2nd Quart., 2015.

[34] A. Sabharwal, P. Schniter, D. Guo, D. W. Bliss, S. Rangarajan, and R. Wichman, "In-band full-duplex wireless: Challenges and opportunities," *IEEE J. Sel. Areas Commun.*, vol. 32, no. 9, pp. 1637–1652, Sep. 2014.

[35] J. I. Choi, M. Jain, K. Srinivasan, P. Levis, and S. Katti, "Achieving single channel, full duplex wireless communication," in *Proc. 16th Annu. Int. Conf. Mobile Comput. Netw. (MobiCom)*, 2010, pp. 1–12. [Online]. Available: http://doi.acm.org/10.1145/1859995.1859997

[36] M. Duarte and A. Sabharwal, "Full-duplex wireless communications using off-the-shelf radios: Feasibility and first results," in *Proc. 44th Asilomar Conf. Signals, Syst. Comput. (ASILOMAR)*, Nov. 2010, pp. 1558–1562.

[37] J. Liu, N. Kato, J. Ma, and N. Kadowaki, "Device-to-device communication in LTE-advanced networks: A survey," *IEEE Commun. Surveys Tuts.*, vol. 17, no. 4, pp. 1923–1940, 4th Quart., 2014.

[38] K. S. Ali, H. ElSawy, and M.-S. Alouini, "On mode selection and power control for uplink D2D communication in cellular networks," in *Proc. IEEE Int. Conf. Commun. Workshops (ICC)*, Jun. 2015, pp. 620–626.

KONPAL SHAUKAT ALI (S'14) received the B.S. degree in electrical engineering from the Lahore University of Management Sciences, Lahore, Pakistan, in 2012, and the M.S. degree in electrical engineering from the King Abdullah University of Science and Technology, Thuwal, Saudi Arabia, in 2014, where she is currently pursuing the Ph.D. degree in electrical engineering. Her research interests include stochastic geometry, modeling of wireless networks, and device to device communication.

HESHAM ELSAWY (S'10–M'14) received the B.Sc. degree from Assiut University, Asyut, Egypt, in 2006, the M.Sc. degree from the Arab Academy for Science, Technology & Maritime Transport, Cairo, Egypt, in 2009, and the Ph.D. degree from the University of Manitoba, Winnipeg, MB, Canada, in 2014, all in electrical engineering. From 2006 to 2010, he was with the National Telecommunication Institute, Egypt, where he conducted professional training both at the national and international levels, as well as research on network planning. From 2010 to 2014, he was with TRTech, Winnipeg, MB, Canada, as a Student Researcher. He is currently a Post-Doctoral Fellow with the Computer, Electrical, and Mathematical Sciences and Engineering Division, King Abdullah University of Science and Technology, Saudi Arabia, and an Adjunct Faculty with the School of Computer Science & Engineering, York University, Canada. His research interests include the statistical modeling of wireless networks, stochastic geometry, and queuing analysis for wireless communication networks. For his academic excellence, he received several academic awards, including the NSERC Industrial Postgraduate Scholarship from 2010 to 2013, the TRTech Graduate Students Fellowship from 2010 to 2014, and the best paper award in the ICC 2015 workshop on small cells and 5G networks. He was recognized as an Exemplary Reviewer by the IEEE Transactions of communication in 2015 and 2016.

ANAS CHAABAN (S'09–M'14) received the Maîtrise ès Sciences degree in electronics from Lebanese University, Lebanon, in 2006, the M.Sc. degree in communications technology from the University of Ulm, Germany, in 2009, and the Dr.-Ing. (Ph.D.) degree in electrical engineering and information technology from Ruhr-University Bochum, Germany, in 2013. From 2008 to 2009, he was with the Daimler AG Research Group on Machine Vision, Ulm, Germany. He was a Research Assistant with the Emmy-Noether Research Group on Wireless Networks, University of Ulm, from 2009 to 2011, which relocated to Ruhr-University Bochum in 2011. He was a Post-Doctoral Researcher with Ruhr-University Bochum from 2013 to 2014. He joined the King Abdullah University of Science and Technology as a Post-Doctoral Researcher in 2015. His research interests include information theory and wireless communications.

MOHAMED-SLIM ALOUINI (S'94–M'98–SM'03–F'09) was born in Tunis, Tunisia. He received the Ph.D. degree in electrical engineering from the California Institute of Technology, Pasadena, CA, USA, in 1998. He was a Faculty Member with the University of Minnesota, Minneapolis, MN, USA, then with the Texas A&M University at Qatar, Doha, Qatar, before joining the King Abdullah University of Science and Technology, Thuwal, Saudi Arabia, as a Professor of electrical engineering in 2009. His current research interests include the modeling, design, and performance analysis of wireless communication systems.