Frontiers of Information Technology & Electronic Engineering
www.jzus.zju.edu.cn; engineering.cae.cn; www.springerlink.com
ISSN 2095-9184 (print); ISSN 2095-9230 (online)
E-mail: jzus@zju.edu.cn

Review:

Multiple-antenna techniques in nonorthogonal multiple access: a review*

Fei-yan TIAN, Xiao-ming CHEN‡

College of Information Science and Electronic Engineering, Zhejiang University, Hangzhou 310027, China

E-mail: tian_feiyan@zju.edu.cn; chen_xiaoming@zju.edu.cn

Received Aug. 8, 2019; Revision accepted Dec. 12, 2019; Crosschecked Dec. 12, 2019

Abstract: As a promising physical layer technique, nonorthogonal multiple access (NOMA) can admit multiple users over the same space-time resource block, and thus improve the spectral efficiency and increase the number of access users. Specifically, NOMA provides a feasible solution to massive Internet of Things (IoT) in 5G and beyond-5G wireless networks over a limited radio spectrum. However, severe co-channel interference and high implementation complexity hinder its application in practical systems. To solve these problems, multiple-antenna techniques have been widely used in NOMA systems by exploiting the benefits of spatial degrees of freedom. This study provides a comprehensive review of various multiple-antenna techniques in NOMA systems, with an emphasis on spatial interference cancellation and complexity reduction. In particular, we provide a detailed investigation on multiple-antenna techniques in two-user, multiuser, massive connectivity, and heterogeneous NOMA systems. Finally, future research directions and challenges are identified.

Key words: Nonorthogonal multiple access; Multiple-antenna technique; B5G; Internet of Things
https://doi.org/10.1631/FITEE.1900405 **CLC number:** TN929.5

Abbreviations

5G	Fifth generation
ADC	Analog-to-digital converter
AF	Amplify and forward
B5G	Beyond 5G
BS	Base station
CDI	Channel direction information
CDMA	Code division multiple access
CE	Channel estimation
CF	Compute and forward
CR	Cognitive radio
CS	Compressive sensing
CSI	Channel state information
DF	Decode and forward
DoF	Degree of freedom
DPC	Dirty paper coding
EH	Energy harvesting
FD	Full duplex
FDD	Frequency division duplex
FRAB	Finite resolution analog beamforming
Gbps	Gigabits per second
HD	Half duplex
IoT	Internet of Things
LOS	Line of sight
LTE-A	Long-term evolution-advanced
MF	Matched filtering
MIMO	Multiple-input multiple-output
MISO	Multiple-input single-ouput
MMSE	Minimum mean square error

‡ Corresponding author

* Project supported by the National Natural Science Foundation of China (No. 61871344), the Zhejiang Provincial Natural Science Foundation of China (No. LR20F010002), the National Science and Technology Major Project of China (No. 2018ZX03001017-002), and the National Key R&D Program of China (No. 2018YFB1801104)

ORCID: Fei-yan TIAN, http://orcid.org/0000-0001-8242-2802; Xiao-ming CHEN, http://orcid.org/0000-0002-1818-2135

© Zhejiang University and Springer-Verlag GmbH Germany, part of Springer Nature 2019

mmWave	Millimeter wave
MUD	Multiuser detection
NLOS	Non-line of sight
NOMA	Nonorthogonal multiple access
OMA	Orthogonal multiple access
PLS	Physical layer security
PS	Power splitting
QoS	Quality of service
RF	Radio frequency
SIC	Successive interference cancellation
SINR	Signal-to-interference-plus-noise ratio
SNR	Signal-to-noise ratio
SOCP	Second-order cone programming
SWIPT	Simultaneous wireless information and power transfer
TAS	Transmitting antenna selection
TDD	Time division duplex
TDMA	Time division multiple access
TS	Time switching
UAD	User activity detection
UE	User equipment
ZF	Zero forcing

1 Introduction

With the fast development of the Internet of Things (IoT), a massive number of IoT devices are poised to access the wireless networks for achieving advanced applications, such as smart city, smart traffic, and smart medicine (Xu LD et al., 2014; Zanella et al., 2014; Catarinucci et al., 2015). The growth rate of IoT devices is tremendous, and it is predicted that in the next decade, at least 100 billion devices will connect to various wireless networks (Chen X , 2019). As shown in Fig. 1, the IoT network is evolving to the Internet of everything. In this context, the biggest wireless network in the world, e.g., the fifth-generation (5G) or even beyond-5G (B5G) wireless network, takes the cellular IoT as its core technique to support massive connections (Palattella et al., 2016; Chen SZ and Kang, 2018; Wan D et al., 2018). For example, the 5G narrowband IoT (NB-IoT) is required to admit 50 000 devices in a cell (TR, 2015). With the explosive growth of the number of IoT devices, the next-generation cellular IoT should realize ultradense access. In this context, it is desirable to adopt a spectral-efficient multiple access technique to support massive connections over limited radio spectrum. However, the currently used orthogonal multiple access (OMA) techniques, e.g., time division multiple access (TDMA), frequency division multiple access, and code division multiple access (CDMA), exclusively allocate one resource block to one user equipment (UE) to avoid possible multiuser interference but with low spectral efficiency. Hence, it is an extremely challenging task for the OMA techniques to simultaneously admit a massive number of UEs with limited radio spectrum. To this end, nonorthogonal multiple access (NOMA) is proposed as an enabling multiple access technique for 5G (Dai LL et al., 2015; Ding et al., 2017d; Liu YW et al., 2017; Ding et al., 2018). In Figs. 2 and 3, we illustrate the differences between OMA and NOMA techniques and show the performance gains of NOMA system on the achievable sum rate and the required total transmit power relative to the OMA system.

In general, NOMA is a multiple access technique, in which multiple UEs are allowed to share the same time-frequency resource block (Ding et al., 2014; Timotheou and Krikidis, 2015; Vaezi et al., 2019). As a result, the receiver receives mixed signals coming from multiple UEs. In other words, NOMA inevitably gives rise to inter-user interference, which decreases the signal quality at the receivers. To facilitate multiuser detection (MUD) at the receiver's end, the received signals

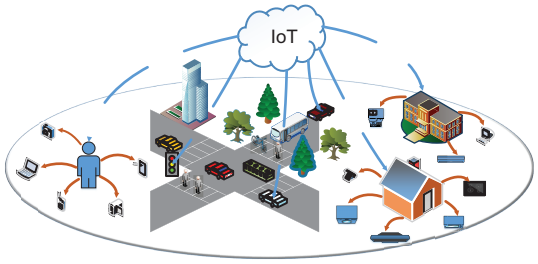

Fig. 1 An illustration of IoT

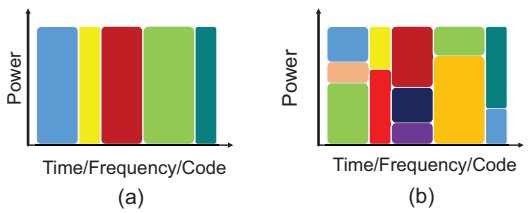

Fig. 2 The comparison between OMA (a) and NOMA (b) techniques

should be easily separated in a certain domain. Thus, according to the domains for signal separation, NOMA can be divided into two categories, namely, power-domain NOMA (PD-NOMA) and code-domain NOMA (CD-NOMA). As the names imply, PD-NOMA exploits the differences in channel gains of UEs for signal detection (Islam et al., 2017; Shin et al., 2017b), while CD-NOMA leverages coding/spreading over multiple carriers to achieve multiuser multiplexing (Peng et al., 2017; Wei F and Chen, 2017). Moltafet et al. (2018a) compared the characteristics and advantages of the two typical NOMA techniques and found that PD-NOMA can be easily combined with the other 5G candidate techniques, e.g., massive multiple-input multiple-output (MIMO) and millimeter wave (mmWave) approaches, to significantly improve the system performance. Thus, PD-NOMA has invited considerable research from various aspects. In this context, this study focuses on the analysis and discussion of PD-NOMA (for convenience, we call it NOMA in the following sections).

Generally, NOMA enables spectral-efficient multiple access by combining superposition coding performed at the transmitter and successive interference cancellation (SIC) at the receiver end (Zhang NB et al., 2016; Zhu JY et al., 2017). To be specific, transmit power, as the weighted coefficient of superposition coding, can be optimized to coordinate the inter-user interference, and SIC is carried out to mitigate the interference from users with weaker channel gains. Therefore, various power allocation schemes have been designed based on different performance metrics, such as sum rate (Yang ZH et al., 2017), weighted sum rate (Liu F and Petrova, 2018), and energy efficiency (Fang et al., 2017). As is well known, the sum rate is commonly involved in the design of multiple access systems. The problem of power allocation using weighted sum rate maximization has been previously investigated (Hu et al., 2017), and it has been found that the problem is convex only in some special cases. Again, Hanif et al. (2016) showed that even for the problem of maximizing the sum rate, power allocation is nontrivial, and the optimal solution exists only when the rate constraints meet a given condition. Thus, it is difficult to provide a general power allocation scheme for multiuser NOMA systems from the perspective of maximizing the sum rate. To this end, power allocation for maximizing the minimum rate has been studied previously (Choi, 2016a). Fortunately, such a problem of power allocation can be transformed as a linear programming problem for a given minimum rate requirement. Thus, based on a bisection search on the minimum rate requirement, it is possible to derive the optimal solution of power allocation. Specifically, this power allocation scheme can guarantee the user data rate fairness to some extent. Similarly, power allocation with the goal of minimizing the total power consumption subject to constraints on the minimum rate is equivalent to a linear programming problem, for which obtaining the optimal solution is easy (Wang H et al., 2017). Moreover, considering the concept of green communication, a power allocation scheme from the viewpoint of maximizing the energy efficiency has been designed (Zeng et al.,

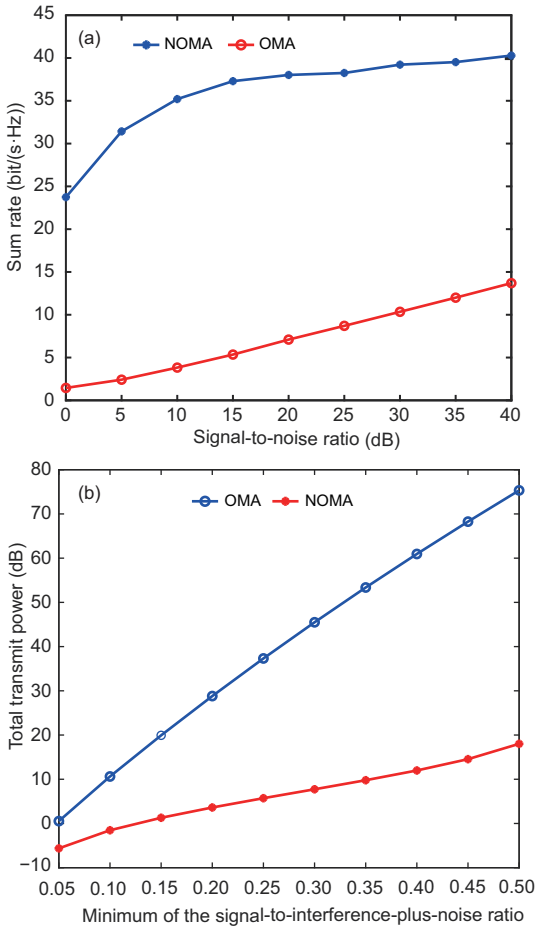

Fig. 3 Performance gains of the NOMA system relative to the OMA system: (a) signal-to-noise ratio; (b) minimum of the signal-to-interference-plus-noise ratio

2018). It is worth pointing out that power allocation can only coordinate, but not mitigate, the inter-user interference. Especially in the scenario of massive access, even with power allocation, there exists strong residual inter-user interference after SIC. In other words, power allocation has limited capability in improving the performance of NOMA systems.

In addition to the system performance, the implementation complexity is an important metric for the design of NOMA systems. In general, the complexity of NOMA is caused mainly by the SIC at the receivers. In the context of massive access, the order of SIC is very large, resulting in high computational complexity. Especially for the receiver with the strongest channel gain, it should cancel the inter-user interference from all the other receivers, and thus the complexity of SIC might be prohibitive. To reduce the computational complexity, it is usual to partition UEs into multiple clusters, with each cluster containing a small number of UEs (Tsai and Wei, 2018; Yang ZH et al., 2018; Zheng et al., 2018). If SIC is only performed within a cluster, then the computational complexity can be significantly reduced. It is intuitive that the criterion of user clustering has a great impact on both the performance and the complexity (Chen C et al., 2017). However, since user clustering is in general an integer programming problem, optimal performance can be achieved only by exhaustive searching, which has a prohibitive computational complexity. Thus, it is difficult to design an optimal user clustering scheme with low complexity. To achieve a balance between system performance and computational complexity, the number of UEs in each cluster is usually fixed at two. Thus, user clustering is reduced to user pairing (Liang et al., 2017). Kang and Kim (2018) provided two simple user pairing schemes. Specifically, the two schemes partition UEs into two groups, according to the sort of channel gains. The first scheme pairs up the strong user in the first group with the strong user in the second group, while the second scheme links the strong user in the first group and the weak user in the second group in the same pair. User pairing has lower complexity compared to user clustering, but it is also difficult to design an optimal user pairing scheme. Moreover, both user clustering and user pairing cause intercluster or interpair interference, which cannot be mitigated by SIC (Ali et al., 2016b). In other words, user clustering reduces the complexity but with performance degradation.

To guarantee the performance of NOMA with user clustering, it is desirable to effectively mitigate the intercluster interference by some means (Ali et al., 2016a; Chen ZY et al., 2016a; Liu ZX et al., 2017). It is intuitive that if the clusters occupy orthogonal resources, intercluster interference can be avoided. Multiple-carrier modulation has been applied in the NOMA system, and a carrier was allocated to only a cluster (Wei ZQ et al., 2017). Thus, intercluster interference is completely cancelled by a simple method. However, multiple-carrier transmission may decrease the spectral efficiency, and it also limits the number of UEs. Since multiple antennas constitute a fundamental characteristic of future 5G and B5G wireless networks (Ali et al., 2017), it is natural to adopt multiple-antenna techniques to combat intercluster interference; thus, all clusters can share the time-frequency resources. In fact, it has been proved that the multiple-antenna technique has a strong capability for mitigating interference by making use of a spatial beam (Zhang J and Andrews, 2010; Chen X and Yuen, 2014; Hosseini et al., 2014). Hence, multiple-antenna NOMA has been widely recognized as a feasible and powerful way of mitigating the intercluster interference and thereby improving the spectral efficiency (Sun Q et al., 2015; Nguyen VD et al., 2017b; Liu L et al., 2019). Nevertheless, multi-antenna arrays may also be coupled mutually among elements; this influence should be considered if the communication quality and spatial beamforming ability need to be further improved (Li and Zhang, 2007; Zhu LF et al., 2015; Mei and Wu, 2018). In general, there are two categories of multiple-antenna NOMA. The first category of multiple-antenna NOMA scheme assigns a spatial beam to each UE, and SIC is performed within a cluster (Jia et al., 2019). Through spatial beamforming, partial or even complete inter-user interference can be cancelled. For instance, if zero-forcing (ZF) beamforming is adopted at the base station (BS) in the scenario of downlink communications, both the intercluster interference and intracluster interference are reduced. As a result, it is likely to achieve high spectral efficiency. Note that this category of multiple-antenna NOMA scheme requires the designing of a spatial beam for each UE. If the number of UEs is large, the computational

complexity of spatial beams might be prohibitive. Moreover, the number of BS antennas limits the number of spatial beams; thus, it is difficult to support massive access based on the first category of multiple-antenna NOMA scheme. For the second category of multiple-antenna NOMA scheme, UEs in a cluster share the same spatial beam, and SIC is carried out within a cluster (Chen XM et al., 2019). It is intuitive that such a multiple-antenna NOMA scheme is capable of decreasing the required number of spatial beams and thus reduces the complexity of beam design. More importantly, it is able to admit a large number of UEs with a finite number of spatial beams. In other words, the second category of multiple-antenna NOMA scheme can effectively improve the spectral efficiency and support massive access over limited radio spectrum. However, since spatial beamforming cannot mitigate the intracluster interference, the second scheme may suffer from severe co-channel interference. Generally, the first scheme can mitigate the co-channel interference but subject to a constraint on the number of UEs, while the second scheme can support massive access albeit with high co-channel interference. Figs. 4 and 5 show the characteristics of these two categories.

To exploit the benefits of multiple-antenna techniques for NOMA, the multiple-antenna transmitter should have partial channel state information

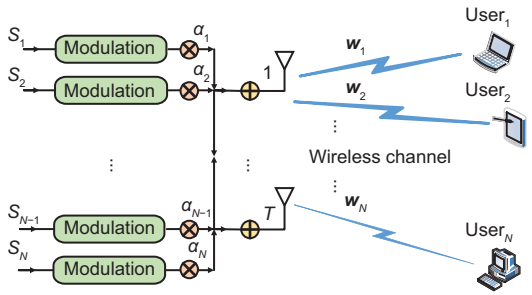

Fig. 4 Characteristic of the first multiple-antenna NOMA scheme

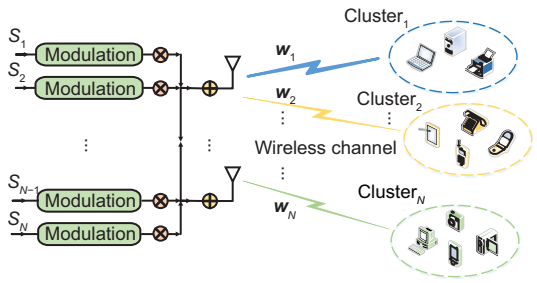

Fig. 5 Characteristic of the second multiple-antenna NOMA scheme

(CSI). Specifically, the multiple-antenna transmitter designs the spatial beams for interference cancellation according to available CSI (Chen C et al., 2017; Cui et al., 2018c; Cheng et al., 2018). The more accurate the available CSI, the lower is the suffered interference. For instance, if the multiple-antenna transmitter has full CSI, it is possible to cancel the co-channel interference for the first multiple-antenna NOMA scheme and mitigate the intercluster interference for the second multiple-antenna NOMA scheme. However, it is not a trivial task to obtain full CSI at the transmitter. In general, the transmitter can obtain only partial CSI in two ways. First, if the system operates in the frequency division duplex (FDD) mode, UEs quantize the estimated CSI and then convey it to the transmitter (Yang Q et al., 2017). Due to the limited capacity of the feedback link, UEs can convey only partial CSI. Second, if the system operates in the time division duplex (TDD) mode, UEs send pilot sequences over the uplink channels, and then the transmitter obtains the CSI about the uplink channels through channel estimation (CE) (Chen XM et al., 2018c). Due to channel reciprocity in the TDD mode, the CSI about the uplink channels can be used for the downlink transmission. Similarly, the CSI at the transmitter is imperfect due to CE error. Thus, the multiple-antenna transmitter only has partial CSI. In other words, there exists uncertainty about the multiple-antenna transmitters. As a result, there is residual co-channel interference at UEs inevitably even with spatial interference cancellation. It is clear that the performance of the multiple-antenna NOMA systems is dependent on the CSI accuracy. In order to improve the overall performance, it is necessary to design an efficient CSI acquisition method, which can obtain accurate CSI with limited resources, according to the characteristics of the multiple-antenna NOMA systems. Meanwhile, the spatial beam should be constructed based on partial CSI. In general, there are two kinds of beam design methods in the case of partial CSI. Specifically, the first method designs the spatial beams based on the available CSI directly (Chen XM et al., 2017a). Since there exists CSI uncertainty at the multiple-antenna transmitter, the first method cannot guarantee the performance of spatial beamforming, especially when the CSI accuracy is low. To solve this problem, the second method designs the beams with the

principle of maximizing the worst performance according to the CSI uncertainty, namely, robust beamforming (Tian et al., 2018). Thus, the second beam design method can provide a performance guarantee under all CSI conditions. However, compared to the first method, the second method has a higher computational complexity and consumes more wireless resources for guaranteeing the performance in the worst case.

Generally, the multiple-antenna techniques are capable of enhancing the performance of the NOMA systems by making use of the spatial degrees of freedom (DoFs). However, the DoFs of multiple-antenna systems are limited by the number of antennas. Thus, in order to admit more UEs, there should be enough DoFs. For instance, if there are a few UEs, it is possible to achieve the performance requirement by deploying a finite number of antennas. As the number of UEs increases, UEs should be grouped into several clusters for reducing the required number of beams. Furthermore, in the scenario of massive access, e.g., IoT, it is difficult for the traditional multiple-antenna systems with a finite number of antennas to satisfy the performance requirements of a massive number of UEs. In this context, it is necessary to deploy a large-scale antenna array to significantly increase the spatial DoFs. Thus, the multiple-antenna techniques can always provide a solution for the design of efficient and reliable NOMA systems.

Multiple-antenna NOMA, due to its promising performance metrics, has received considerable attention. So far, there have been numerous multiple-antenna techniques for enabling NOMA under different conditions. Although this topic has been studied extensively, numerous theoretical and technical issues still remain open. Moreover, we have not seen a complete analysis to reveal multiple-antenna techniques in NOMA systems in terms of fundamental results, recent advances, and future trends. For example, Wei ZQ et al. (2016b) and Dai et al. (2018) presented a primary investigation of NOMA techniques in 5G wireless networks, but they did not focus on the potentials and challenges of multiple-antenna NOMA. In fact, as discussed earlier, the multiple-antenna technique has a great impact on the performance of NOMA systems. In this study, we investigate state-of-the-art research results from the perspective of multiple-antenna NOMA and point out potential directions and challenges for future works.

2 Fundamental techniques of multiple-antenna NOMA systems

The principle behind multiple-antenna NOMA is that the use of multiple-antenna techniques mitigates the severe co-channel interference caused by nonorthogonal transmission of multiple UEs. Especially in PD-NOMA systems, received signals are required to have distinct power levels for facilitating signal detection with a mixed signal. In general, if UEs have distinct access distances, it is easy to separate the received signals in the power domain. As seen in Fig. 6, UE_1, close to the BS, may receive a stronger signal compared to UE_2. Thus, UE_1 first detects the signal of UE_2, subtracts the interference caused by UE_2's signal, and then recovers its own signal. However, in practice, UEs might be close to each other, e.g., in hotspot areas. Although power control can be used to distinguish UEs to some extent, its capacity is limited due to the power constraint. Fortunately, by adjusting the spatial beams of multiple-antenna NOMA systems, it is possible to effectively separate UEs even at the same access distance. Especially, if the UE with a long access distance requires a higher channel gain for achieving better performance relative to the UE with a short access distance, spatial beamforming can be applied to enhance the channel gain of the farther UE. Generally, benefits of the multiple-antenna technique for enabling NOMA can be summarized as follows:

1. Mitigate co-channel interference.
2. Enhance the desired channel gain.
3. Separate UEs in the spatial domain.
4. Admit more UEs and networks.
5. Simplify the detection at UEs.

To achieve these benefits, it is desired to apply the multiple-antenna techniques according to

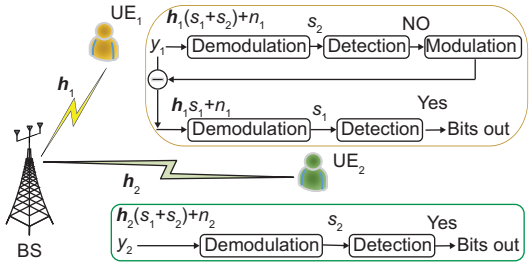

Fig. 6 A simple two-UE system model

the characteristics of the NOMA systems. In particular, the fundamental techniques of NOMA systems should be fully combined with the spatial DoFs offered by the multiple-antenna systems. As a simple example, UEs are partitioned into several clusters in the spatial domain, so as to decrease the computational complexity of SIC at UEs. In what follows, we introduce the fundamental techniques of multiple-antenna NOMA systems in detail.

2.1 Channel state information acquisition

CSI availability at the transmitters is a precondition for the design of multiple-antenna NOMA techniques. In particular, user scheduling, beam design, resource allocation, and signal detection need the CSI (Chen XM et al., 2010; Chen XM and Zhang, 2010; Shi et al., 2018; Yalcin et al., 2019). In the ideal case, the transmitters may obtain full CSI by some means (Dai et al., 2017; Seo and Sung, 2018; Xiao ZY et al., 2018). If the transmitter has full CSI, it is possible to completely cancel the co-channel interference by using spatial interference mitigation techniques, i.e., ZF beamforming (Ding et al., 2016b). The resource consumption, i.e., transmit power, also can be minimized if the system is resource constrained (Choi, 2015). However, it is not a trivial task to acquire full CSI at the transmitter in multiple-antenna systems. This is because the estimation and conveyance of multiple-dimensional CSI may consume considerable wireless resources, which is unacceptable in practical systems. As a simple example, the BS in long-term evolution-advanced (LTE-A) systems is equipped with four or eight antennas. Then, the CE at UEs requires four or eight orthogonal training symbols, and conveyance of the estimated CSI in terms of a complex vector to the BS is hardly achieved. To solve it, the CSI is usually quantized with a codebook as a finite number of code words (Love et al., 2008). The instantaneous CSI is represented by an optimal code word in some sense, and thus the CSI feedback just needs a few bits. In this way, the BS only obtains partial CSI, and the CSI accuracy at the BS is determined by the spatial resolution, namely, the ratio of the number of feedback bits to the number of antennas (Dai et al., 2008). Unlike simple multiple-antenna systems, the NOMA systems suffer from severe inter-user interference. Thus, NOMA systems require accurate CSI for designing efficient transmission schemes. As a result, the amount of CSI feedback should be increased linearly proportionally to the number of UEs. Due to the limited capacity of the feedback link, the total number of feedback bits is constrained. It is clear that the equal allocation of feedback bits may lead to a low feedback efficiency over the fading channels. Especially in NOMA systems, UEs experience different levels of inter-user interference. Specifically, the UE with the strongest channel gain can experience mitigation of all inter-user interference but would interfere with all the other UEs, while the UE with the weakest channel gain would suffer all inter-user interference but may not interfere with any other UEs. In other words, UEs require different CSI accuracies from the perspective of optimizing the overall performance (Xi and Zhou, 2016). The net throughput based on limited CSI feedback, defined as the difference between the downlink information transmission rate and the uplink CSI feedback rate, is taken as the optimization objective for distributing the feedback bits (Chen XM et al., 2012). Since the net throughput is a complicated function of the feedback bits, it is difficult to obtain a closed-form solution. To reduce the computational complexity of feedback bit allocation, Chen XM et al. (2017a) proposed an algorithm with the goal of minimizing the total inter-user interference. Note that feedback bit allocation is an integer programming problem indeed, and the optimal solution is usually derived through numerical searching. Thus, low-complexity algorithms provide only suboptimal solutions.

The CSI acquisition method based quantized feedback is usually applied in FDD systems. In fact, several systems are operated in the TDD mode. For instance, massive MIMO systems in 5G networks are commonly suggested to work in the TDD mode (Lu et al., 2014; Elijah et al., 2016) because the required number of feedback bits for meeting a given CSI accuracy might be prohibitive in the FDD mode. Fortunately, due to the characteristic of channel reciprocity in the TDD mode, the CSI about the uplink channels can be used as that of the downlink channels. As such, the BS can directly obtain the downlink CSI by estimating the uplink channels based on the training method. Specifically, UEs send orthogonal training sequences over the uplink channels, and resulting CSI accuracies are determined by the transmit energy of training sequences,

namely, the product of transmit power and the duration (Hoydis et al., 2013). Therefore, for a given duration of the training sequences, it is possible to adjust CSI accuracies through power control at UEs according to the characteristic of the NOMA systems. Although the use of orthogonal training sequences is beneficial to simplify the CE, it also leads to long training sequences. In the scenario of a large number of UEs, the duration of orthogonal training sequences may occupy a very high proportion of a time slot. In the worst case, the length of the training sequence exceeds the duration of channel coherence time. As a result, the estimated CSI is outdated and thus invalid (Mi et al., 2017). To resolve this problem, nonorthogonal CE is introduced into multiuser NOMA systems (Chen XM and Jia, 2018). To be specific, multiple UEs share the same training sequence, but training sequences are orthogonal to each other. Thus, the required number of training sequences is significantly decreased, and the length of the training sequences is shortened. It is worth pointing out that sharing of the training sequence produces co-channel interference, resulting in decrease of the CSI accuracy. Moreover, CSI accuracies related to UEs sharing the same training sequence interrelate. Hence, it makes sense to coordinate the transmit powers of UEs for improving the overall performance. The order of SIC in multiuser NOMA systems has been exploited in Xiao L et al. (2018) to optimize the transmit powers at UEs. In general, the length of the training sequence and the CSI accuracy are contradictory. It is necessary to achieve a balance between the length of the training sequence and the CSI accuracy by adjusting the number of UEs sharing the same training sequence. However, since the optimization of the number of UEs is an integer optimization problem, it is usual to design a heuristic algorithm for determining the number of UEs.

The above two CSI acquisition methods should perform CE at the beginning of each time slot. In some scenarios, channel states are fast time-varying, and thus it is hard to obtain instantaneous CSI in time. Moreover, the two methods consume considerable wireless resources during CSI estimation and feedback, which might be unbearable in resource-constrained NOMA systems. To alleviate the impact of time-varying channel fading on the CSI acquisition, several works have proposed the use of channel statistical information to design multiple-antenna NOMA schemes (Wei ZQ et al., 2016a; Chitti et al., 2017; Wang XS et al., 2018). In general, channel statistical information can be obtained by averaging numerous channel realizations. The major advantage of channel statistical information is that it remains constant over a relatively long time period, which sharply reduces the overhead for CSI acquisition. The ergodic capacity of MIMO NOMA systems based on channel statistical information has been analyzed in Sun Q et al. (2015). It has been shown that even with channel statistical information, the NOMA scheme performs much better than the OMA scheme. Then, channel statistical information was applied to design the FDD-based large-scale MIMO NOMA systems. As analyzed earlier, massive MIMO systems are usually operated in the TDD mode for ease of CSI acquisition. However, in some scenarios, the massive MIMO system should adopt the FDD mode. Fortunately, channel statistical information can be used to address the problem of CSI acquisition in FDD-based massive MIMO NOMA systems, although it may lead to a performance loss compared to the case with instantaneous CSI (Choi, 2016b). Furthermore, if the massive MIMO NOMA system works at mmWave band, the characteristic of sparsity of the channel due to the limited number of scatters can be explored to construct the statistical channel information (Wang BC et al., 2017; Zhang D et al., 2017). Specifically, the statistical channel information can be obtained based on the information of the angle of arrival, which avoids the averaging operations over a large number of channel realizations.

Additionally, in some extreme scenarios, any kind of CSI is difficultly acquired at the transmitter. In this case, it is only possible to design the open-loop multiple-antenna NOMA scheme (Chraiti et al., 2018). For instance, Ding et al. (2016a) have advocated the use of ZF detection at UEs to mitigate co-channel interference. However, in order to cancel all the interference, the number of antennas at the UE should be greater than the number of antennas at the BS, which is unpractical in real systems. In general, there are not enough spatial DoFs at the UE to cancel the interference in multiuser MIMO NOMA systems. We provide a summary of the CSI acquisition methods for multiple-antenna NOMA systems in Table 1.

Table 1 CSI acquisition methods

CSI type	CSI acquisition method
Quantized instantaneous CSI	CSI is quantized by a codebook and then the index of the quantization code word is fed back (Chen XM et al., 2017a)
Estimated instantaneous CSI	Pilot sequences are sent over the uplink channels and then CSI is estimated at the BS directly (Chen XM and Jia, 2018)
Statistical CSI	A large number of channel realizations are averaged (Choi, 2016b)
Angular domain CSI	Angular domain channel information is conveyed (Wang BC et al., 2017)
No CSI	No action is required (Ding et al., 2016a)

2.2 User clustering

User clustering is used to achieve a balance between system performance and computational complexity in multiuser scenarios (Ding et al., 2016c). In multiple-antenna NOMA systems, UEs are partitioned into several clusters in the spatial domain by exploiting the spatial DoF (Zeng et al., 2017; Ding et al., 2019). UEs in a cluster share the same spatial beam, and SIC is only performed within a cluster. In general, user clustering can be divided into two subproblems; the first one is the number of clusters, and the second one is UEs in a cluster. Given the total number of UEs, the number of clusters influences the intercluster interference, and the number of UEs in a cluster determines the intracluster interference. Therefore, user clustering affects the intracluster and intercluster interference directly and thus has a great impact on the performance of NOMA systems. In order to optimize the performance, user clustering should be dynamically adjusted with UE mobility and channel varying (Ali et al., 2016a). However, it is not a trivial task to design a dynamic user clustering scheme. This is because user clustering is an integer programming problem, and the optimal numerical searching method leads to a high computational complexity. To this end, it is usual to design some suboptimal user clustering schemes according to the spatial characteristics of multiple-antenna NOMA systems, i.e., spatial correlation and spatial orthogonality (Zaw et al., 2017). Hence, it is easy to mitigate the intercluster interference by using multiple-antenna spatial beamforming techniques.

As mentioned earlier, to design the user clustering scheme, the first step is to determine the number of clusters. In general, the number of clusters is limited by the number of BS antennas. This is because the number of independent spatial beams must be not greater than the number of BS antennas. Given the upper bound of the number of clusters, it is still difficult to determine the optimal number of clusters. The simplest and commonly used method is to allow two UEs in a cluster, namely, user pairing (Chen X et al., 2018; Zhu LP et al., 2018). The two-user clustering method has low intracluster interference and can mitigate the intercluster interference through spatial beamforming. In other words, this method can effectively cancel the co-channel interference caused by nonorthogonal transmission; thus, it is suitably adopted in the case of interference limited regimes. However, in the regions of low and medium transmit power, the co-channel interference might not be dominant (Chen XM et al., 2017a). As such, the two-user clustering method leads to low spectral efficiency. Moreover, it limits the number of admissible UEs due to a finite number of transmit beams. In this context, Tsai and Wei (2018) have proposed a dynamic mode selection according to channel conditions and system parameters. Specifically, the number of clusters, namely, transmission mode, can be adjusted for maximizing the sum rate of multiple-antenna NOMA systems.

Given the number of clusters, the second step in the design of the user clustering scheme is to select UEs for each cluster, which is a user scheduling problem indeed. It is intuitive that user scheduling in multiple-antenna systems is performed based on the available CSI. According to the CSI types, the scheduling of users forming the clusters can be categorized as follows.

2.2.1 Instantaneous CSI

If the instantaneous CSI is available at the multiple-antenna transmitter, it is possible to schedule UEs according to the concerned performance metrics directly. A user clustering scheme from the perspective of maximizing the throughput of

the worst user has been proposed in Liu YW et al. (2016), so as to improve the user's data rate fairness in the NOMA system.

2.2.2 Statistical CSI

As discussed in Section 2.1, mmWave channels in general are sparse in the angular domain. Cui et al. (2018b) proposed a user clustering algorithm based on machine learning to maximize the sum rate and meet the quality of service (QoS) requirement at UEs with weak channel gains. Their statistical channel model of mmWave systems is constructed based on information on the angles of departure, which thus avoids a large number of channel realizations and sharply reduces the complexity.

2.2.3 Channel gain information

Kim et al. (2013, 2015) proposed two algorithms for user clustering in multiple-antenna NOMA systems, both of which formulate an optimization problem to minimize the total power consumption and process the user clustering by the ranking approach. In the channel gain correlation-based approach, the algorithm is conducted in an ascending order. On the contrary, the scheme is carried out in a descending order in the channel gain difference-based approach.

2.2.4 User position information

Due to the limited nature of spectrum, Zhang XK et al. (2017) have investigated user clustering based on user locations for NOMA visible light communication multicell networks in order to reduce the interference, as high precise positioning is feasible. The CSI is available at the transceiver side and the residual interference during the SIC in NOMA is also taken into account in the optimization problem. Table 2 shows a summary of the user clustering schemes with different CSI types.

2.3 Superposition coding

Superposition coding is the key for realizing efficient nonorthogonal transmission. Generally, superposition coding is equivalent to the weighted sum of UEs' signals, and thus they can share the same spectrum resource (Wang JH et al., 2017; Moltafet et al., 2018a; Sun YS et al., 2018). For power-domain NOMA, the transmit powers are usually used as the weighted coefficients (Lei L et al., 2016; Xu P and Cumanan, 2017). Consequently, the mixed signals can be separated in the power domain at the receiver. In multiple-antenna NOMA systems, rather than the power domain, the spatial domain can be exploited to improve the performance (Nguyen VD et al., 2017b; Liu L et al., 2019). In other words, superposition coding in multiple-antenna NOMA systems consists of two parts: (1) spatial beamforming that exploits the spatial DoF for mitigating the intercluster interference; (2) power allocation that adjusts the weighted coefficients for coordinating the intracluster interference. In what follows, we introduce the two parts of superposition coding in multiple-antenna NOMA systems.

2.3.1 Spatial beamforming

Nonorthogonal transmission leads to severe co-channel interference, which may decrease the spectrum efficiency. Fortunately, multiple-antenna techniques can be used to mitigate the interference. The simplest multiple-antenna technique for superposition coding is antenna selection. By selecting one antenna or multiple antennas with high channel gains but weak co-channel interference, this method can effectively improve the performance under conditions of limited resources and cost. Intuitively, the global optimal solution of antenna selection may require an exhaustive search over all possible antenna combinations, the complexity of which would become unacceptable as the numbers of antennas at both the BS and UEs are large. To reduce the complexity, Yu et al. (2017b) proposed two simple antenna selection algorithms, namely, max-min-max antenna selection and max-max-max antenna selection. They can asymptotically approach the optimal performance. Moreover, as we know, each antenna in multiple-antenna systems usually requires one dedicated radio frequency (RF) chain, and RF components may consume up to 70% of the total transceiver energy consumption. Therefore, we may face problems of high transceiver complexity and energy consumption in mmWave massive MIMO systems. Antenna selection-based NOMA scheme can address these challenges. First, the use of NOMA makes multiple UEs share one beam, which decreases the required number of RF chains. Then, the antenna selection technique can further improve the energy efficiency. Wang BC et al. (2017) proposed an algorithm, combining NOMA and beamspace

Table 2 User clustering schemes with different CSI types

CSI type	Definition of the user clustering scheme
Instantaneous CSI	Throughput (Liu YW et al., 2016): the maximum throughput of the worst user
Statistical CSI	Sum rate (Cui et al., 2018b): the maximum sum rate of meeting QoS requirements of weak users
Channel gain information	Power consumption (Kim et al., 2015): the minimum total power consumption with channel gain correlation-based scheme
	Power consumption (Kim et al., 2013): the minimum total power consumption with channel gain difference-based scheme
User position information	User rate (Zhang XK et al., 2017): the achievable user rate under user QoS constraint

antenna selection, to significantly reduce the number of required RF chains in mmWave massive MIMO-NOMA systems. In addition, the study provided a joint power allocation scheme that optimized not only the intrabeam power but also the interbeam power.

Antenna selection cannot fully exploit the spatial DoF of multiple-antenna NOMA systems, resulting in an obvious performance loss. A more powerful multiple-antenna technique is the use of spatial beamforming (Fig. 7), according to channel conditions and system parameters (Alavi et al., 2017; Shin et al., 2017a). In general, there are two categories of beamforming design methods for multiple-antenna NOMA systems. The first method designs the spatial beams based on a certain principle directly. For example, the multiple-antenna NOMA broadcast channels are not generally degraded, of which the capacity region can only be attained by using dirty paper coding (DPC) (Chen ZY et al., 2016b). However, DPC is difficult to implement in practice due to its nonlinearity and prohibitive complexity. In this context, NOMA schemes with linear beamforming for nondegraded broadcast channels have gained researchers' attention. In multiple-antenna NOMA systems, there are the following linear beamforming schemes.

1. Random beamforming

A beamforming matrix is stored at both sides of transmitter and receiver in advance. Receivers feed back the effective signal-to-noise ratio (SNR) of the optimal beamforming vector based on instantaneous CSI. Then, the transmitter assigns beamforming vectors for all receivers according to the feedback information. Thus, random beamforming has low complexity and does not require the transmitter to obtain CSI. Results from Ding et al. (2017b) show that random beamforming can yield significant

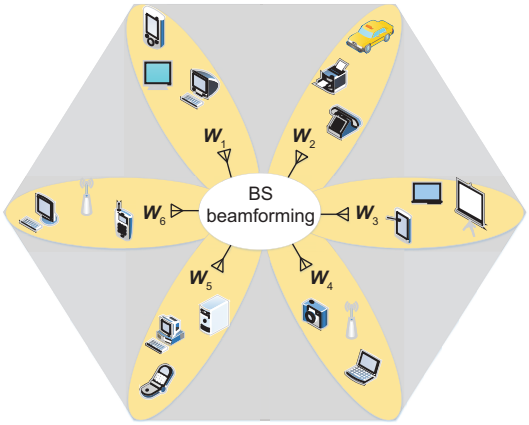

Fig. 7 An illustration of the beamforming model

performance gain over conventional MIMO OMA schemes.

2. Analog beamforming

Analog beamforming does not alter the amplitude of a signal but modifies its phase only. Ding et al. (2017a) proposed a new NOMA transmission algorithm that exploits the feature of finite resolution analog beamforming (FRAB). Authors have designed a single FRAB-based beamformer, which is shared by multiple UEs and which has proved its excellent performance.

3. Digital beamforming

Common beamforming schemes include mainly matched filtering (MF), ZF, and minimum mean square error (MMSE) (Cai et al., 2017; Hu et al., 2017). The main idea of MF is to maximize the signal gain of the target UE, but it does not consider the interference between different UEs. It is only applicable to scenarios with low channel correlation and is more suitable for scenarios with a large number of BS antennas. ZF, in contrast, is dedicated to eliminating interference between different UEs and ignores the effects of noise, and the total achievable rate of the system is lower than in the MF scheme.

The goal of MMSE is to minimize the mean square error between the received signal and the transmitted signal. It is more common in scenes with fewer users due to its high complexity.

The second method designs the spatial beams by solving an optimization problem based on a given performance metric. Hanif et al. (2016) have optimized the transmit beams by maximizing the sum rate in the downlink multiple-antenna NOMA system. The study approximates a nonconvex and intractable problem with a minorization-maximization algorithm, solves a series of second-order cone programming (SOCP) problems under decidability constraints, and then obtains the complex beamforming vectors by developing an iterative algorithm. Moreover, the minimization of total power consumption is a commonly used optimization objective for beam design in multiple-antenna NOMA systems (Wang H et al., 2018; Bai et al., 2019). Note that in some extreme cases, the above two beam design methods might be inapplicable. For example, if the channels have correlated CSI, the above methods cannot design different beamforming vectors for UEs. To this end, Mitra and Bhatia (2017) proposed a new hybrid beamforming scheme for downlink NOMA channels using a closed loop adaptive Chebyshev predistorter, based on singular value decomposition, in IoT applications. It works even when all the channel matrices are correlated. The summary of various multiple-antenna superposition coding schemes is given in Fig. 8 and Table 3.

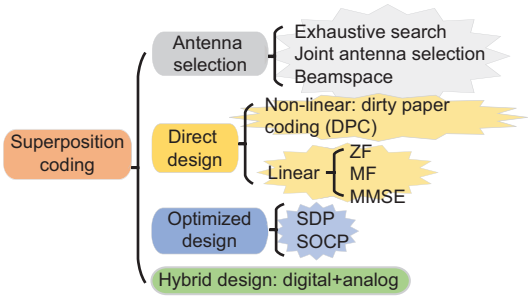

Fig. 8 An illustration of superposition coding schemes

2.3.2 Power allocation

Power allocation is a commonly used method for power-domain NOMA to separate the mixed signals. In multiple-antenna NOMA systems, power allocation is used mainly for coordinating the intracluster interference (Amin et al., 2018; Celik et al., 2019; Tong et al., 2019). Intuitively, the simplest method is the fixed proportional power allocation. For instance, the UE with the weakest channel gain is allocated to the highest power. Thus, UEs may achieve fair data rates. However, fixed proportional power allocation does not strictly guarantee system performance due to its inflexibility.

To satisfy the performance requirements, transmit power can be allocated based on a given metric. Zeng et al. (2018) proposed an energy-efficient power allocation scheme for multiple-antenna NOMA systems under the premise of ensuring QoS and meeting the minimum rate requirement for each UE. If it is not feasible, a low-complexity user admission protocol has been proposed, which admits users one by one following the ascending order of required power for satisfying the QoS requirements. Similarly, the optimization indicator proposed in Zhang HJ et al. (2018) is also energy efficient, but the difference is that researchers discuss the performance of the proposed scheme for an NOMA heterogeneous network with both perfect and imperfect CSI. If the CSI is perfect, fractional transmit power allocation is exploited to further allocate power among users, which is superior to equal power allocation. On the contrary, when BSs only have the estimated value of CSI, an outage probability requirement should be considered as constraints for resource scheduling. In this case, a unique optimal solution can be found by the gradient-assisted binary search algorithm. Results demonstrate that the system's energy efficiency deteriorates when the channel gain estimation error variance increases. The difference between the two resource allocation methods is shown in Table 4.

2.4 SIC in NOMA

SIC is a key technique in NOMA systems for combating co-channel interference and thus improving the performance. The basic principle of SIC is to gradually reduce the interference from other users. The SIC detector performs data decision on multiuser signals one by one at the receiver and judges whether it belongs to itself. If not, the multiple access interference caused by this user signal is subtracted. In NOMA, the transmitter allocates the power for users. The SIC operates in the order of the magnitude of the signal power, and the higher power signal operates first because the strongest one

Table 3 Summary of the beamforming strategy

Strategy	Main feature
Random beamforming	Transmitter does not need to know CSI (Ding et al., 2017b)
Analog beamforming	Antenna selecion with finite resolution (Ding et al., 2017a)
Digital beamforming	MF: aims to maximize the signal gain of target user without considering interference ZF: aims to eliminate interference while ignoring the effects of noise MMSE: aims to minimize the mean square error
Optimized beamforming	Joint optimization: beamforming and other problems are simultaneously considered (Nguyen VD et al., 2017a; Wang H et al., 2018) Robust beamforming: aims to improve the worst-case sum rate when CSI is imperfect (Zhang Q et al., 2016)
Hybrid beamforming	Traditional digital method and optimization scheme are combined (Cai et al., 2017; Hu et al., 2017)

Table 4 Two typical power allocation methods

Method	Characteristics
Fixed proportional allocation	Simple but inflexible
Optimization allocation	CSI is perfect (Zeng et al., 2018): aiming at ensuring QoS or the minimum rate requirement CSI is imperfect (Zhang HJ et al., 2018): the outage probability is also considered

is more likely to capture it. This process is repeated until the desired data are found. Finally, a user will cancel the signal of the user whose power is stronger to obtain its own signal, and the remaining signals of the latter users still exist as interference. The process of SIC is shown in Fig. 9.

If the SIC is perfect, the interference is only the above residual interference. Receivers must observe a certain signal power disparity, which is determined mainly by the hardware sensitivity. On the contrary, if the SIC is imperfect, decoding imperfections leave some additional interference at the very end. There are some problems when using imperfect SIC, e.g., an error propagation problem (decoding errors of earlier users will continually affect the decoding of later users). As a result, many robustly joint MUDs are developed.

An imperfect SIC receiver model has also been investigated (Celik et al., 2017), wherein the power disparity, sensitivity constraints, delay tolerance, and residual interference due to detection and estimation errors are considered in building the model.

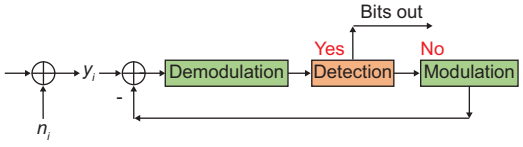

Fig. 9 The process of successive interference cancellation (SIC)

Authors concluded that the NOMA gain is limited by channel gain disparity and SIC characteristics. They include the residual interference after cancellation, uncanceled interference due to affordable number of cancellations, and uncanceled lower rank interference into the generic signal-to-interference-plus-noise ratio (SINR) representation; the latter cluster formulation and power-bandwidth allocation problem are based on this expression. Results show the impacts of constraints and imperfections on achievable performance.

Similarly, other researchers (Chen XM et al., 2019) assumed that the SIC process is not perfect; therefore, they considered the hardware limitation of mobile terminal and found that residual interference from weaker users caused by multiple factors, e.g., coding/modulation-related parameters and error propagation, still exists. Then, based on this imperfect SIC, a linear model is adopted and a joint optimization algorithm from the perspective of minimizing the total power consumption is proposed to maximize the weighted sum rate.

In addition, there is a fault condition that the decoding time might increase due to the mismatch between the user index and its actual place in the cancellation sequence. In other words, the decoding order does not match the user index. In this case, users will not obtain their own signals and will request their messages from the BS, which leads to

an increase in processing time. Thereafter, the time for decoding at the SIC receiver with and without user mismatch is discussed (Manglayev et al., 2017). When users mistakenly decode the signal of adjacent users, the final decoding time for both the nearest user and the farthest user grows.

In the traditional SIC receiver, if any of the previous users fails in decoding, outage for the current user will inevitably occur. However, we hope that the receivers, under the premise that the decoding order is fixed, attempt to decode the latter users' signal by regarding the unsuccessfully decoded users' signal as the interference, in order to improve the outage performance. An advanced multiuser SIC decoding strategy based on statistical CSI in uplink NOMA systems was also proposed (Xia et al., 2018). Different from the traditional SIC receiver, in the proposed scheme, the BS will not stop the detection for the following users when one user fails to decode. Consequently, without introducing any performance loss on the prior decoded users, there will be a significant gain because outage balancing among different users can be achieved. The various modeling schemes for SIC are summarized in Table 5.

3 Two-user NOMA systems

Obviously, introducing multiple antennas makes the NOMA system more complex; not only user clustering and power allocation issues, but also beamforming, needs to be considered. In order to reduce the complexity, the SIC receiver is mostly applied to two-user scenarios. Moreover, in a multiuser system, two users satisfying certain channel conditions are often grouped into one cluster. Therefore, the two-user scenario is the simplest and most basic model, the analysis of the characteristics of which is general.

As seen in Fig. 10, we usually simply think that in two-user systems, one user is near and the other is far away; we also call them strong channel gain user and weak channel gain user. The weak user always has poor outage performance.

Consider a simple case. Assume that the BS knows the channel direction information (CDI) of the strong and weak users and the statistical characteristics of the channel, and that the strong and weak users have the same CDI but the channel statistic characteristics are different (the channel of the weak user has a certain degree of fading relative to that of the strong user). According to the strong user's CDI, ZF is used for beamforming to obtain the received signal vector, and the weak user's received vector is also obtained. At the receiver, the strong user performs SIC after correctly decoding the symbol of the weak user to gain the desired signal. However, the weak user does not have the SIC process, and the strong user's signal is regarded as interference to decode its signal directly. Throughout the process, certain measures can be taken to optimize the system capacity or the outage performance of the weak user.

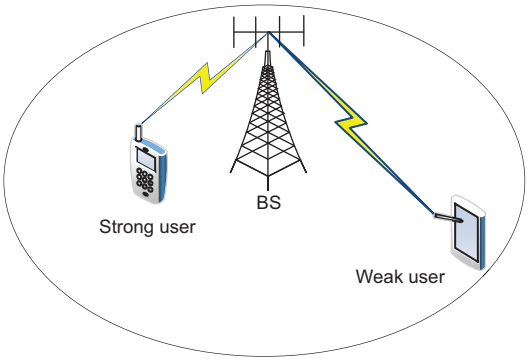

Fig. 10 Model of the two-user NOMA system

In existing NOMA techniques, when users share the same DoF, using SIC allows, at the most, one user to get one DoF, while the remaining users get zero DoF. Therefore, the available resource is not fairly shared among the users. Moreover, power allocation is possible only when the instantaneous or statistical CSI is available at the transmitter. If the CSI is unavailable, there will be some problems, such as allocation of power among users. In this context, Chraiti et al. (2018) introduced an NOMA scheme for a two-user multiple-input single-output (MISO) downlink channel with unknown CSI, which allows the transmitter to communicate with two users simultaneously while keeping the signals perfectly separable at the respective receivers. Authors have broken up the available DoF into fractional DoFs such that all users get a nonzero DoF. Therefore, the proposed scheme that fairly shares a DoF among users and removes the requirement of CSI is significant.

Regarding the power allocation in the two-user systems, using the outage probability and the average rate as the criteria to analyze the performance, a novel dynamic power allocation scheme has been proposed (Yang Z et al., 2016), whereby the power allocation factors dynamically change with the

Table 5 The views and models in various cases

Scenario	Views and models
Perfect SIC	Only the residual interference is left, and the system performance is determined by hardware sensitivity
Imperfect SIC	Additional interference needs to be considered when modeling, such as detection and estimation errors (Celik et al., 2017), the hardware limitation of the mobile terminal (Chen XM et al., 2019), wrong cancellation order (Manglayev et al., 2017), and the outage caused by the decoding failure (Xia et al., 2018)

instantaneous channel gains. The proposed scheme can more flexibly meet various QoS requirements and ensure the individual user rates by constructing two constraints, which means the situation that the weak user is served with a small data rate can be avoided. As a result, different trade-offs between the user's data rate fairness and system throughput are realized, and performance gain is guaranteed.

To perform superposition coding, beamforming is also required. As mentioned earlier, beamforming can be achieved by random, traditional, and optimization methods. For the two-user system, we introduce only the optimized beamforming technique with better performance, and others will not be described again. An optimal precoding scheme for QoS optimization in a two-user multiple-antenna system has been investigated earlier (Chen ZY et al., 2016b), wherein the minimal total transmission power and the optimal precoding vectors have been obtained by considering their Lagrange dual problem under the condition of two given target interference levels. Due to the small number of users, there is less constraint on the individual achievable rate required by each user, and the closed-form solution of the optimization problem can be easily derived. Therefore, the complexity is sharply reduced in two-user systems.

The purpose of studying the two-user system is to simplify the highly complicated scenarios. It deserves to be studied because it can be considered as the basic unit in multiuser or even massive-user systems.

4 Multiuser NOMA systems

In multiple-antenna NOMA systems, user clustering is first done by forming as many clusters as the number of transmitting antennas and groupings of multiple users in each cluster; subsequently, multiple users in each cluster share the spatial dimension and are served in the power domain. Then, the problem of optimal beamforming and power allocation compatible to user scheduling should be solved to maximize the performance with control of interference. Obviously, the complexity of the NOMA system is determined mainly by the complexity of the SIC receiver. Therefore, in the multiuser system, the users are usually clustered. A clustered multiuser NOMA multiple-antenna system model is shown in Fig. 11. If it is downlink, the SIC is implemented in a cluster containing a small number of users. In other words, user clustering is an important issue to achieve a balanced trade-off between the complexity of the SIC decoding technique and the performance of the NOMA system. What needs to be addressed in multiuser systems is the user clustering. The user clustering schemes classified according to the different CSI levels have been introduced in detail earlier; here, we analyze the two scenarios of uplink and downlink.

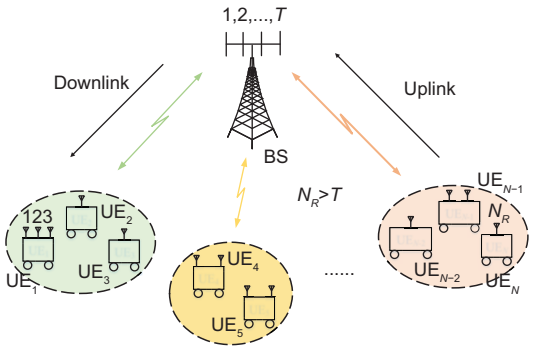

Fig. 11 Model of the multiuser NOMA system

In NOMA downlink, SIC should be implemented at user terminals. When the number of users at each cluster is large, SIC needs high processing power. Moreover, the loss of practical SIC should not be neglected. As a result, grouping only two users into a resource block in NOMA downlink has become more popular, and some problems in

multiuser systems can be solved based on the analysis of two-user systems. In the literature for the NOMA downlink system, user clustering has been studied considering aspects such as the data rate fairness guarantee, minimization of the outage probability, maximization of the sum rate, and minimization of the transmit power. Authors are also researching the user clustering algorithm that considers both the channel correlation and gain difference among users. In addition, the relationship between the number of receiving and transmitting antennas in a multiple-antenna system needs attention. The number of clusters is generally equal to the number of BS transmitting antennas. At this time, a single beam is used by all the users in a cluster, while in the case of more clusters than BS transmitting antennas, multiple clusters may share the same beam. Actually, in a multiple-antenna system, the inter-user interference can be completely eliminated when the number of total receiving antennas is equal to or less than the total number of transmitting antennas in a cell. However, in multiple-antenna NOMA downlink, the number of receiving antennas is more than the number of transmitting antennas; thus, the interference for each user is very high. To minimize the net interference and maximize the system capacity, a low-complexity multiple-antenna NOMA user clustering technique for downlink transmission has been proposed (Ali et al., 2016b), whereby UE receiving antennas were grouped dynamically into a number of clusters equal to or more than the number of BS transmitting antennas. Then, authors combined a new ZF-beamforming technique with this user clustering scheme to cancel the intercluster interference. It can maintain robustness for a wide range of transmitting antennas at the BS and for different cluster sizes. Similarly, another dynamic user clustering problem has been investigated (Liu YW et al., 2016) to realize the different trade-offs of complexity and throughput of the worst user. In that study, authors have studied the issue considering the aspect of data rate fairness.

In the uplink, SIC is at BSs, which have enough processing power to perform SIC. Thus, users do not need to be aware of the modulation and coding schemes used by the other users. Specially, in NOMA uplink, users are also allowed to transmit in a grant-free manner, which can reduce latency significantly. In the single-antenna scene, some suboptimal user pairing, iterative detection schemes, and optimum resource allocation techniques have been already investigated in existing works. In addition to single-antenna applications, user pairing and power allocation plans for NOMA uplink have been studied previously for multiple-antenna wireless networks. Note that the relationship between the number of antennas at each user terminal and the number of antennas at the BS sometimes needs to be considered. However, there is no analytical evaluation of performance in most of the previous studies for optimum user pairing in NOMA uplink. In another study, Sedaghat and Müller (2018) proposed an optimal user pairing for some given suboptimal power allocation schemes in the uplink, which can be implemented using algorithms with polynomial time complexity. This scheme shows how much the performance is improved if the users in the NOMA uplink are paired optimally. If it is a cellular network, a single isolated cell is considered and inter-cell interference is neglected. The BS first divides the users into some clusters and lets each cluster transmit at one of the subcarriers. At the multi-antenna receiver, the signals at each subcarrier are detected by dividing the users into pairs and applying SIC to each pair. If the user terminals had a single antenna, the BS directly divided the users into pairs and detected every pair using SIC. In this case, the optimum pairing method had significant performance gain compared to the random pairing method. Meanwhile, when the case of multi-antenna users is considered, the BS pairs the users based on the CSI and the users need not know the pairing strategy. In this scheme, the ranks of the channels and the number of antennas at users are not required, and the users need only to know their own channels. The proposed technique also outperforms NOMA with signal alignment.

5 Massive NOMA systems

In 5G mobile communication systems, almost everything of society will be connected. At the same time, several serious challenges are posed for 5G and B5G, such as the requirements for millions of connections per square kilometer, single-millisecond end-to-end latency, user experienced data rate in gigabits per second (Gbps), and more severe physical layer security (PLS). Both theory and a large number of system-level simulation results verify that NOMA is

very suitable for 5G and B5G application scenarios of ultralow latency and ultrahigh connectivity, and it is expected to support high spectral efficiency and massive connectivity with scarce spectrum resources for IoT in urban areas. IoT has the characteristic of connecting a very large amount of devices but usually with a small data rate requirement. To address these challenges, future networks will need to meet this special requirement by efficiently using the available bandwidth.

Although the number of total potential users can be quite large, the rate of active users is usually very low, e.g., Fig. 12. Thus, IoT requires a completely different set of techniques to support massive connectivity. These techniques are usually called user activity detection (UAD). Because UAD is generally performed at the BS, it is mainly for the uplink scenario. In current 4G systems, the uplink transmission is scheduled by the BS in a request-grant procedure, with resultant large transmission latency and signaling overhead. This problem is unacceptable for massive connectivity in 5G. Therefore, it is highly expected that users can randomly transmit data in uplink and a BS does not know which users are active, i.e., grant-free transmission, and thus user activity needs to be detected. The uplink grant-free NOMA technique, which has received a lot of attention, has been heavily investigated to reduce control signaling overhead and transmission latency of scheduling procedure (Chen ZL et al., 2018; Du et al., 2018a; Senel and Larsson, 2018). The following paragraphs describe the various UAD schemes in the uplink grant-free NOMA systems.

Typically, the UAD based on compressive sensing (CS) is a promising field to exploit the user activity sparsity due to sporadic communication in IoT. However, the signal detection in the CS-based scheme is usually independently realized in different time slots, while the user activity correlation in different time slots is not considered. Therefore, researchers have considered the more practical scenario that active user sets can be changed in several continuous time slots and have proposed a low-complexity dynamic CS MUD to jointly realize user activity and data detection (Wang BC et al., 2016). Furthermore, they have used the estimated active user set in the current time slot as the initial set to estimate the active user set in the next time slot under the framework. The proposed solution is

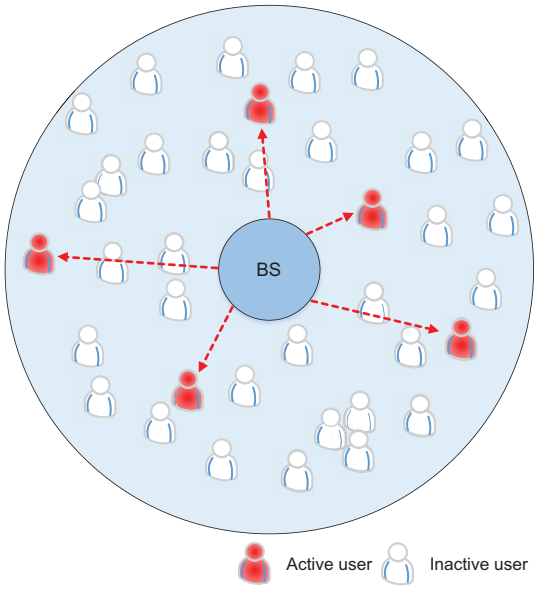

Fig. 12 Model of the massive NOMA system

superior to the conventional CS-based scheme in terms of performance.

As mentioned earlier, to reduce control signaling overhead and latency, the uplink grant-free NOMA systems, in which active users can transmit data at any time slot without a complex request-grant procedure, have been studied. In current wireless systems, the number of active users is usually much smaller than the total number of possible users even during busy hours. This characteristic also applies to massive connectivity in 5G. Thus, the inherent sparsity of active users could be exploited to solve the MUD problem. Meanwhile, the framewise joint sparsity of user activity, i.e., sharing the same locations of nonzero elements during several successive time slots within a frame, can be used to further improve the performance of MUD. Because users generally transmit their information based on a uniform frame structure, the inactivity and activity of users would remain constant over an entire data frame.

In existing works, joint UAD and CE by sending pilots have been carried out to perform MUD (Liu L et al., 2018; Shao et al., 2019). In CS-based works, it is assumed that the channels for all users are known a priori. In fact, most of the users are inactive and the channel information is then outdated, and if the CSI is perfect to all users every time, it means that the BS does not need to ask users whether they are willing to send data or not. Therefore, the method of joint UAD and CE with

the aid of pilot symbol is reasonable. However, framewise joint sparsity, i.e., the temporal correlation characteristic, is not used in some recent works on joint UAD and CE. A novel joint UAD and CE scheme for uplink grant-free NOMA systems has been investigated (Du et al., 2018b), wherein the inherent framewise joint sparsity of the pilot and data phased in the entire frame is used to formulate the joint CE and UAD problem under multiple measurement vector-CS framework. Then, authors proposed the block sparsity adaptive subspace pursuit algorithm without requirement for the user sparsity level as the prior information: it exploits the block-sparse single measurement vector-CS model, which is transferred by multiple measurement vector-CS to improve the performance.

Similarly, joint UAD and data detection using the framewise joint sparsity of user activity has been implemented to perform MUD in recent reports. Compared with joint UAD and CE, it has a better performance. However, in some works, the joint UAD and data detection scheme does not use the prior information of the transmitted discrete symbols, and there still is room for improvements. Therefore, Wei C et al. (2017) have proposed a joint algorithm for uplink grant-free NOMA systems, wherein expectation maximization based on the framewise joint sparsity of user activity has been employed to estimate user activity parameters, while prior information of the transmitted discrete symbols has been used by the approximate message passing technique to compute the posterior means and variances. This scheme shows a much better performance. Table 6 shows several kinds of UAD algorithms.

6 Heterogeneous NOMA systems

It is well known that NOMA can significantly improve the spectrum efficiency of a system, but it does not necessarily meet all the performance metric requirements in certain specific situations. Often encountered are trade-offs between user's data rate fairness and QoS, trade-offs between the sum rate of the system and the minimum rate requirements of weak users, trade-offs between user throughput and outage probability, trade-offs between energy efficiency and capacity of system, and so on. It is obvious that NOMA needs to be combined with existing advanced techniques to achieve greater performance enhancements and optimizations. The following subsections describe several common variations of the multiple-antenna NOMA network.

6.1 Relay NOMA

Cooperative relay is an effective technique to extend coverage areas and overcome channel impairments, such as fading, shadowing, and path loss. The fundamental form of cooperative communications is the dual-hop relaying, in which the source communicates with the destination via the assistance of a relay. The decode-and-forward (DF) and amplify-and-forward (AF) protocols are two basic relay protocols; regeneration and nonregeneration categories are used in the two protocols, respectively (Yuksel and Erkip, 2007; Chen XM et al., 2015c). The function of the regeneration mode is that the relay attempts to first decode the information stream and then retransmit the decoded stream to the destination, which is always used by strong users to improve the performance of weak users. On the contrary, in the non-regeneration mode, the relay simply retransmits a scaled version of its observation to the destination. The two relay processes are shown in Fig. 13.

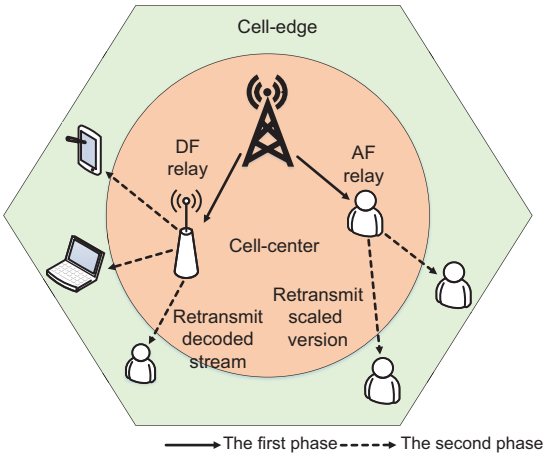

Fig. 13 Model of the cooperative relay NOMA system

Various cooperative relay NOMA schemes have been extensively studied in the existing literature (Zhong and Zhang, 2016; Liu X et al., 2017; Chen XM et al., 2018a). Zhou Y et al. (2018) proposed a dynamic DF cooperative NOMA scheme for downlink transmission, wherein the researchers considered random- and distance-based two-user pairing strategies. Outage probabilities were derived

Table 6 Classification of UAD algorithms

Category	Common practice
Conventional scheme	CS: detection is independently realized in different time slots
	Dynamic CS (Wang BC et al., 2016): the user activity correlation in different time slots is considered
Exploiting framewise joint sparsity of the active user	Joint UAD and CE (Du et al., 2018b): with the aid of pilot symbols
	Joint UAD and data detection (Wei C et al., 2017): the transmitted discrete symbols are prior information

for each user-pairing strategy, from which the diversity order and the sum rate can be obtained. The proposed scheme can achieve a diversity order of two for the weak user without sacrificing spectral efficiency. On the other hand, an AF relay network has been proposed, wherein joint power allocation and relay beamforming design problem have been investigated for this network (Xue et al., 2017). The authors have studied an alternating optimization-based algorithm and transformed the power allocation problem into an SOCP, namely, the relay beamforming problem into a convex linear-fractional programming, to maximize the achievable rate of the destination that has the best channel condition. However, under the assumption that only the statistical CSI is known, both DF and AF protocols were considered in Wan DH et al. (2018), and all nodes in the system were equipped with a single antenna. The authors found that DF relaying always outperforms the AF one in terms of the resulting ergodic sum rate and outage probability in the low-SNR region.

In Li et al. (2018), multiple antennas have been assumed for the BS and all UEs in the half-duplex (HD) cooperative NOMA system. In order to improve the overall throughput of the network, the BS first broadcasts the superposed signals to a multiple-antenna central user; then, the central user acts as a relay node and helps the BS cooperatively relay signals for the cell-edge user using the DF relaying protocol. The achievable rate from the BS to the cell-edge user is maximized under transmit power constraints and achievable rate constraints from the BS to the central users. Also using the DF protocol, the performance analysis in the full-duplex (FD) cooperative NOMA system has been discussed in a parallel study (Zhang L et al., 2017). The difference in data rate fairness between the near user and far user is also taken into account while minimizing the outage probability of the system and maximizing the minimum achievable rate of users.

Thus, how to choose the forwarding protocol and the working mode of the system also becomes a key issue. The study and comparison of the performances of the dual-hop relaying system exploiting the DF and AF strategies, respectively, have been undertaken by deriving the closed-form expressions for outage probability and ergodic capacity (Xiao Y et al., 2018). Meanwhile, the authors studied the adaptive relay forwarding strategy selection between DF and AF NOMA under fixed and optimized user power allocation to obtain better performance. Similarly, a pair of relay selection schemes were considered (Yue et al., 2018b), wherein the closed-form expressions for the exact and asymptotic outage probabilities were derived to evaluate the performance of the two proposed schemes. Especially, the performances of the relays working in the FD or HD mode were compared in this study. The FD-based scheme can obtain a zero diversity order and the HD-based scheme is capable of achieving a diversity order of K (the number of relays). Results demonstrate that the FD-based scheme has better outage performance than the HD-based one in the low-SNR region, but they are both superior to random relay selection schemes. Similar conclusions can be seen in Yue et al. (2018a), in which DF relay was exploited by the near user to help the far user; the energy efficiency was also used as a performance indicator. The study derives that in the delay-limited transmission mode, FD NOMA has higher energy efficiency in the low-SNR region, while in the delay-tolerant transmission mode, HD NOMA is more energy efficient in the high-SNR region.

The above-mentioned cooperative relay NOMA transmissions are all for the downlink scenario, and the uplink scenario is similar. A relay NOMA technique for uplink interference-limited cellular network was proposed by Shin et al. (2017c), wherein the AF protocol was used and the relay was operated in the HD mode. Furthermore, any CSI at the users was

not required and only limited CSI was required at the relays and the BS. With the diversity gain of two, an Alamouti structure of the desired symbol can be generated at the BS without interference in the proposed scheme. Additionally, a buffer-aided relay selection policy for the uplink of NOMA networks has been investigated (Nomikos et al., 2019). The summary is given in Table 7.

6.2 PLS in NOMA

With the popularity of wireless communication in everyday life, immense amounts of private and sensitive data is transmitted over wireless channels. Thus, secrecy issues have become critical due to the unavoidable open nature of the wireless medium. PLS has been widely regarded as a promising cryptographic technique to secure data transmission over wireless channels (Khisti and Wornell, 2010; Chen X and Chen, 2014; Yang N et al., 2015; Chen XM et al., 2017b; Gomez et al., 2019; Nguyen NP et al., 2019). As seen in Fig. 14, PLS issues can be roughly divided into two research directions. First, based on the pioneering concept of the wiretap channels proposed by excellent scholars, the framework for PLS has been widely investigated, whereby the transmitter wishes to send confidential messages to the legitimate receiver in the presence of external eavesdropper (Chen XM et al., 2015b; Chen J et al., 2016; Zhang Y et al., 2016b). On the contrary, the other research direction on PLS does not focus on the secrecy issue against external eavesdroppers. Instead, confidential messages are sent to the intended users while maintaining ignorance at other users' end in the same network (Chen XM et al., 2016, 2018b; Chen X and Zhang, 2017).

The first PLS issue in NOMA systems has been extensively studied in recent years. The optimal designs of decoding order, transmission rates, and power allocation for a secure NOMA system have been studied in detail (He et al., 2017), whereby the transmitter sends confidential messages to multiple users in the existence of an external eavesdropper, and the instantaneous CSI of the eavesdropper is unknown. In this case, the transmit power is minimized subject to the secrecy outage and QoS constraints, and the minimum confidential information rate among users is maximized subject to the secrecy outage and

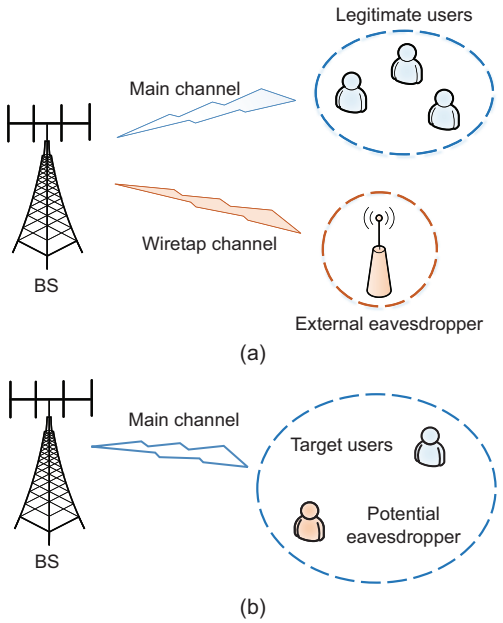

Fig. 14 Illustration of two PLS issues: (a) external eavesdropper; (b) internal eavesdropper

transmit power constraints by using an iterative algorithm. Moreover, in the presence of a multiple-antenna eavesdropper, a novel secrecy beamforming scheme that exploits artificial noise was proposed for the secure multiple-antenna NOMA transmission (Lv L et al., 2018a), wherein imperfect SIC was assumed and the limitations of conventional artificial noise-aided scheme were overcome; i.e., artificial noise was generated in the null space of the main channel such that only the channel of the eavesdropper was degraded. To evaluate the secrecy performance, the closed-form expressions for the secrecy outage probability and the secrecy diversity order were derived. Furthermore, Lei HJ et al. (2018) took a more comprehensive situation into account for the secure NOMA system, whereby both noncolluding and colluding eavesdroppers were considered. Therefore, the exact and asymptotic closed-form expressions for the secrecy outage probability were derived when noncolluding and colluding schemes are used for the eavesdroppers, respectively. Besides, the max-min transmitting antenna selection scheme has been investigated to improve the secrecy performance.

Instead of considering the external eavesdropper, other works (Ding et al., 2017c) have investigated the secure NOMA network with mixed multicast and unicast traffic, where the transmitter wants to send the confidential message to only the

Table 7 Comparison of different relay modes

Relay protocol	System mode	Performance comparison
AF (Xue et al., 2017)	HD (Shin et al., 2017c) FD (Xue et al., 2017)	The DF protocol always outperforms the AF protocol (Wan DH et al., 2018), and the FD-based scheme always outperforms the HD-based one (Yue et al., 2018b).
DF (Zhou Y et al., 2018)	HD (Li et al., 2018) FD (Zhang L et al., 2017)	The forwarding strategy and mode selection are given in a few studies (Xiao et al., 2018b; Yue et al., 2018a, 2018b)

target user and keep all other users ignorant of it. Beamforming and power allocation coefficients were jointly designed to ensure the performance of the system. Similarly, a cellular multiple-antenna NOMA secure transmission system was studied, whereby the central user was an entrusted user and the cell-edge user was regarded as a potential eavesdropper that may eavesdrop on the messages to the central user (Li et al., 2017). The secure beamforming and power allocation design optimization problem of maximizing the sum achievable secrecy rate of central users subject to the transmit power constraint at the BS and transmission rate constraint at cell-edge users has been analyzed.

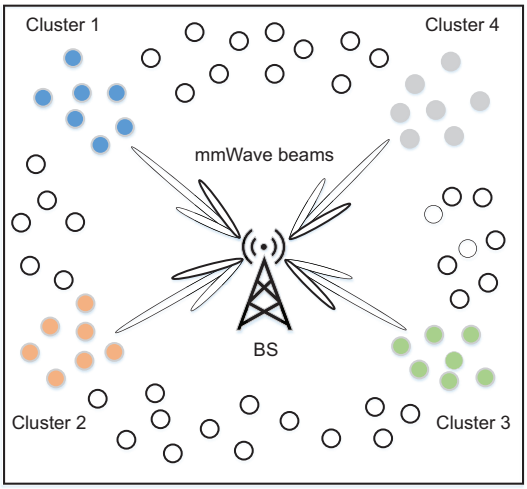

Fig. 15 Model of the multiple-antenna mmWave NOMA system

6.3 mmWave NOMA

Due to the high demand for bandwidth caused by significantly increased data rates, mmWave is naturally applied to NOMA systems (Fig. 15). Accordingly, mmWave-NOMA transmission has a huge potential for satisfying the requirements of IoT (Lv T et al., 2018). The key features of mmWave propagation are the high directionality with severe propagation path loss, low penetration coefficients, and high signal attenuation (Zhang D et al., 2017; Hu et al., 2019; Liu PL et al., 2019; Xiao ZY et al., 2019; Zeng et al., 2019). Consequently, the typical mmWave channel model, which contains a line-of-sight (LOS) path and several non-line-of-sight (NLOS) paths, is generally analyzed. Obviously, the path loss of NLOS exponents is much larger than that of the LOS exponent; therefore, the effect of the LOS path is dominant if such a path exists. Otherwise, the dominant path is one of the NLOS paths.

Compared with the conventional low-frequency multiple-antenna system, the additional RF hardware constraints existing in mmWave systems is another feature of mmWave transmission. In this case, digital baseband beamforming becomes impossible. Meanwhile, analog beamforming is inconvenient since the modulus of the elements in the analog beamforming vectors is constrained to a constant. Therefore, hybrid analog and digital beamforming for mmWave systems has been studied in recent works (Sohrabi and Yu, 2016; Wu QQ et al., 2017; Zhang SQ et al., 2017). However, the designs of the beamforming matrices are generally based on perfect CSI, which is difficult to achieve in practice. In other words, conventional beamforming schemes require that all users provide their accurate CSI to the BS, leading to the system overhead and latency inevitably increasing. Random beamforming is an effective approach to reduce the feedback overhead. Two random beamforming approaches have been proposed for mmWave-NOMA systems in order to avoid the requirement of the BS for all the users' CSI and reduce the system overhead (Ding et al., 2017b). Their performance has been analyzed in terms of sum rates and outage probability, and the simulation results demonstrated that the proposed schemes are indeed capable of achieving significant performance gains. Consequently, in mmWave channels, random beamforming is always used to achieve a better sum rate performance with appropriate user

clustering and power allocation schemes.

In the literature on user clustering and the power allocation technique, random beamforming is also used to improve the performance of mmWave-NOMA systems. For instance, based on the distance between the BS and users, a novel user-pairing technique designed for cellular machine-to-machine communication using mmWave-NOMA systems for IoT applications has been proposed (Lv T et al., 2018). The study adopts the single-path mmWave channel simplified model; i.e., it considered only the LOS path, and the outage probability and the sum rate were used as the performance metrics. Due to the high directionality of the mmWave system, the proposed mmWave-NOMA transmission can easily achieve massive connectivity in cellular communications. Furthermore, these pairing schemes that exploit random beamforming do not require the BS know the CSI of all users, thus naturally reducing the system overhead and latency. Similarly, an optimal user scheduling and power allocation strategy for mmWave-NOMA systems with the aid of random beamforming has been investigated (Cui et al., 2018a), in which the problem of maximizing the sum rate subject to the users' QoS requirements is tackled without introducing high system overhead or heavy computational complexity.

6.4 Cognitive radio and NOMA

Conventional power-domain NOMA allocates more power to the weaker users, which ensures user's data rate fairness but cannot strictly guarantee users' QoS targets. However, cognitive radio (CR) NOMA can meet the QoS requirements of some or even all users (Chen XM et al., 2014a; Lv L et al., 2016; Yu et al., 2017a; Xu L et al., 2018; Wang XY et al., 2019). As shown in Fig. 16, in a CR network, the user with the poorer channel condition is usually regarded as the primary user; i.e., this user will be served with sufficient power to strictly satisfy its data rate and QoS requirements. The concept of primary network is similar. The other network is called the secondary network. Two networks interfere with each other. As an excellent technique, CR has drawn significant attention due to its high spectrum efficiency (Chen XM and Chen, 2013; Chen XM and Yuen, 2013; Wu YN and Chen, 2016). Both CR and NOMA techniques are promising to improve the spectrum efficiency and user

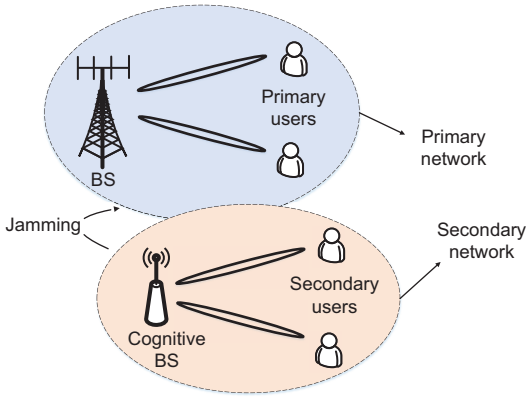

Fig. 16 Model of the CR-NOMA system

connectivity. Combining the CR with multiple-antenna NOMA, not only can IoT devices meet their target QoS needs, but also an additional network can be allowed access, which has tremendous potential in increasing the throughput of the entire system and the number of users to be served. However, some challenges still need to be addressed. For instance, the mutual interference between the primary and the secondary networks may be more severe due to the nonorthogonal nature of NOMA, which will decrease the spectrum efficiency of CR. Many efforts have been made to facilitate the application and investigate the performance of CR-NOMA. In fact, the core idea of CR-NOMA is to design a power allocation strategy to cater to the predefined QoS needs of users.

Additionally, in CR networks, there are three operation paradigms, namely, interweave mode, underlay mode, and overlay mode (Zhou FH et al., 2018b). So far, most of the research work on CR-NOMA has been done on the latter two modes.

The situation of the interweave mode is relatively complicated. The transmission process is divided into two time slots, called spectrum sensing slot and data transmission slot. During the first slot, spectrum sensing is performed to detect whether the frequency bands are occupied by primary users. In the data transmission slot, when detecting that the primary users are inactive, the secondary network can access the frequency bands of the primary users and serve multiple secondary users by using NOMA. Otherwise, spectrum sensing continues to be performed to find the available frequency bands. In this mode, false decisions caused by fading may occur in practice, so interference should be considered.

Under the underlay mode, the secondary

network coexists with the primary network on the premise that the interference caused by the secondary users is tolerable to primary users. The outage probability of the cooperative underlay CR-NOMA DF relaying network has been investigated with primary interference and imperfect CSI (Arzykulov et al., 2019). The authors found the optimal power allocation factors to satisfy the outage probability and data rate fairness and proved that the proposed system model is superior to cooperative OMA. Ding et al. (2016b) have studied the application of MIMO techniques to CR-NOMA under the underlay mode. The impact of different power allocation strategies, namely, fixed and CR-inspired power allocation, on the performance of MIMO CR-NOMA was also discussed. Results demonstrated that combining the existing MIMO-NOMA with the power allocation strategy inspired by the CR network can achieve a significant performance gain in terms of outage probability. Zhang Y et al. (2016a) proposed an efficient algorithm to optimize the performance of an underlay downlink CR-NOMA system from the perspective of energy efficiency. The unauthorized secondary user can be admitted into the spectrum on the condition that it will not cause performance degradation to primary users. The problem of maximizing energy efficiency subject to the individual QoS constraint for each primary user is analyzed. By introducing the multiple antenna technique, the proposed NOMA framework outperforms the conventional NOMA.

In the overlay mode, the secondary network provides cooperation for the primary network to gain access to the frequency band. A novel overlay spectrum-sharing framework has been proposed to enhance the spectrum utilization for single-antenna CR-NOMA systems (Lv L et al., 2018b), whereby one secondary user is scheduled to help forward the primary signals as well as convey its own signals by using NOMA. A reliability-oriented secondary user scheduling scheme and a data rate fairness-oriented secondary user scheduling scheme were compared in terms of outage probability and user data rate fairness. Furthermore, an NOMA cooperative transmission scheme for overlay CR network was investigated (Lv L et al., 2017), wherein the secondary transmitter serves as a relay and helps transmit the primary and secondary messages simultaneously by exploiting the NOMA principle. The proposed scheme outperforms the conventional NOMA in outage probability and system throughput for both primary and secondary networks. Similarly, a cooperative spectrum sharing CR-NOMA system (Kader and Shin, 2016) was considered, in which the secondary transmitter acting as a DF relay for the primary system is allowed to transmit its own signal by superposing on the space-time block coded primary signal. The ergodic capacity of the proposed protocol has been shown to be better than that of the general superposition coding-based overlay CR scheme. Table 8 summarizes the main features of the three modes.

6.5 Simultaneous wireless information and power transfer in NOMA

In cellular multiple-antenna NOMA networks, as we all know, the spectrum efficiency of both cell-edge and cell-center users can be enhanced significantly by exploiting NOMA. However, the user throughput fairness is a critical issue since the data rate of cell-edge users is often lower than that of cell-center users. As reported earlier (Choi, 2015), if cell-edge users want to reach the data rate comparable to that of cell-center users, the power allocation coefficient of cell-center users needs to be close to zero. This is definitely not feasible because the QoS requirements of the cell-center users cannot be met. An efficient way to solve the user data rate fairness issue while ensuring QoS of users may be the cooperative relay NOMA scheme mentioned earlier. The sum rate of the system can be significantly improved when using cooperative relay NOMA, but how cell-center users that act as relay fairly consume their energy to process their own signals and forward the signals of cell-edge users becomes a new question.

Since RF signals can carry both information and energy, simultaneous wireless information and power transfer (SWIPT) is introduced as a promising solution (Liu JX et al., 2008; Varshney, 2008; Chen XM et al., 2013, 2014b; Qi and Chen, 2019; Wu W et al., 2019), whereby cell-center users can scavenge energy from the environment and use the harvested energy to power their relaying operation. Due to the incorporation of SWIPT into NOMA, the limitation of energy storage at the relay nodes is alleviated. Relay nodes can weigh the trade-off well between the information receiving process for themselves and information forwarding process for others. In other words, the relay nodes finally forward the

Table 8 Three operation paradigms of CR

Operation paradigm (Zhou FH et al., 2018b)	Characteristics
Interweave mode	Relatively complicated, false decisions caused by fading may occur
Underlay mode	Secondary network coexists with the primary network under some premise (Ding et al., 2016b; Zhang Y et al., 2016a; Arzykulov et al., 2019)
Overlay mode	Secondary network provides cooperation for the primary network (Kader and Shin, 2016; Lv L et al., 2017, 2018b)

information by using the harvested energy only and do not consume energy from their battery. Additionally, the time switching (TS) and power splitting (PS) energy harvesting receivers are two practical receiver architectures in SWIPT (Chen XM et al., 2015a). The detailed models are shown in Fig. 17. Many efforts have been directed toward the SWIPT NOMA systems using the TS or PS scheme.

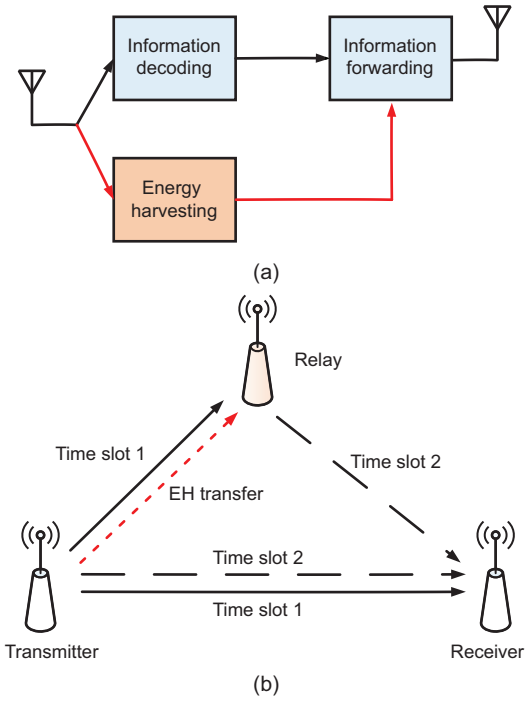

Fig. 17 Comparison of the two energy-harvesting strategies: (a) power splitting architecture; (b) time switching architecture

Xu YQ et al. (2017) investigated a new cooperative SWIPT MISO-NOMA protocol. The strong user (or cell-center user) acts as an energy harvesting (EH) relay to help the weak user (or cell-edge user) improve the communication reliability. The strong user performs SWIPT by exploiting the PS scheme; i.e., the signal received at the strong user is split into two parts, one for information decoding and the other for EH. Particularly, the joint design of beamforming and PS was considered and the problem of maximizing the data rate of the strong user subject to the QoS requirement of the weak user was analyzed. Results demonstrated that the cooperative SWIPT NOMA protocol outperforms the existing cooperative transmission protocols. On the other hand, the TS protocol has been used in a DF relay SWIPT NOMA system (Nguyen TS et al., 2018). The whole block time is divided into EH time and information transmission time, with the proportion being called the TS ratio. Finally, the expressions of outage performance and the delay-limited throughput are derived to evaluate the robustness of the system. Simulation results proved that the system performance is affected by the placement of the relay node. Likewise, a two-user MISO-NOMA system has been studied (Do et al., 2018), wherein cooperative schemes that use hybrid SWIPT and the transmitting antenna selection (TAS) technique were proposed to improve the outage performance and data rate fairness of the cell-edge user. In these schemes, the cell-center user acts as a DF relay to assist the cell-edge user, and its relaying operation is powered by a hybrid TS/PS SWIPT protocol. The proposed schemes with different TAS criteria achieve better outage performance compared to conventional OMA and noncooperative NOMA systems.

Differently, Dai et al. (2019) investigated the integration of SWIPT into mmWave massive MIMO-NOMA systems. They also used hybrid analog and digital precoding to significantly reduce the number of required RF chains without obvious performance loss. In the proposed systems, each user can extract both information and energy from the received RF signals by applying SWIPT with the PS receiver. Furthermore, user grouping was performed based on the correlation of users' equivalent channels, and power allocation for mmWave massive MIMO-NOMA and the PS factors for SWIPT were jointly optimized to maximize the achievable sum

rate of the system. The proposed scheme is capable of achieving higher spectrum and energy efficiency compared with MIMO-NOMA.

In addition, Zhou FH et al. (2018a) combined the multiple-antenna CR-NOMA network with SWIPT to support massive power-limited battery-driven devices. Researchers have applied a more practical nonlinear EH model instead of an ideal linear EH model. An artificial noise-aided cooperative jamming scheme was also proposed to improve the security of the primary network. Both the problem of artificial noise-aided beamforming design subject to the practical secrecy rate and EH constraints and the problem of minimizing the transmission power have been tackled. Results showed that this scheme can achieve significant performance. Table 9 lists several application scenarios for each EH model.

7 Future research directions and challenges

Multiple-antenna techniques have been proven to be effective solutions to enhance the performance and decrease the complexity of NOMA by theoretical analysis, simulation verification, and trial measurement. However, there are many challenging issues that need to be addressed in future research. In what follows, we introduce these research directions and the corresponding challenges.

7.1 Access protocol design

The currently adopted multiple-access protocol, namely, grant-based random access, is designed for OMA systems with a finite number of UEs. Specifically, the grant-based random access protocol requires four times negotiations between the BS and the UE to build the connection based on orthogonal preamble sequences. As mentioned herein, multiple-antenna NOMA will be widely used in 5G and B5G wireless networks with a massive number of UEs. As a result, the grant-based random access protocol may lead to high overhead and high latency. Hence, it is necessary to design a new access protocol for multiple-antenna NOMA systems. Recently, a grant-free random access protocol has been proposed, which allows UEs to simultaneously access wireless networks without a grant. Yet, the grant-free random access protocol has a high complexity. Therefore, one should design low-complexity grant-free random access protocols according to the characteristics of multiple-antenna NOMA systems.

7.2 Low-cost multiple-antenna techniques

In 5G and B5G wireless networks, the BS will be equipped with a large number of antennas. Hence, the use of massive MIMO NOMA scheme can significantly improve the spectrum efficiency and admit a large number of UEs. Usually, each antenna corresponds to an RF chain, but the cost of massive MIMO NOMA is prohibitive. Moreover, the accuracy of the analog-to-digital converter (ADC) of each RF chain has a great impact on the cost. In order to reduce the cost for practical applications, on the one hand, it is necessary to adopt a few number of RF chains. On the other hand, low-resolution ADC for each RF chain should be used. Yet, the combination of a few number of RF chains and low-resolution ADC may result in performance degradation. In fact, NOMA can decrease the required number of RF chains but needs a high-resolution ADC. Thereby, it makes sense to design low-cost multiple-antenna techniques based on the requirements of NOMA.

7.3 Weak coupling antenna structure

The multiple-antenna technique is widely used in 5G and B5G wireless systems. However, the influence of mutual coupling among the antenna elements and arrays in a multiantenna terminal is usually underestimated. Especially for mmWave communication and IoT with massive connectivity, the channel state is sensitive to the antenna structure. The influence of antenna structural error, including the element and array positioning accuracy and the mechanical distortion resulting from the operating environment, should be taken into account in performance analysis and optimization of the system. To tackle this problem, some electromechanical coupling technologies can be adopted (Haupt and Rahmat-Samii, 2015; Wang CS et al., 2017, 2018). Recently, some scholars have experimentally investigated the impact of mutual coupling on data throughput by using an LTE module as the test bed and an anechoic chamber for over-the-air measurement. Even so, research on weak coupling antenna structure is still worth exploring.

Table 9 Application of the EH mode

System framework	EH mode (Chen XM et al., 2015a)	Definition
DF relay	PS	Data rate (Xu YQ et al., 2017): the maximum data rate of strong user subject to the QoS of weak user
	TS	Robustness (Nguyen TS et al., 2018): good outage performance and delay-limited throughput
	Hybrid TS/PS	Outage performance (Do et al., 2018): the data rate fairness of cell-edge user
mmWave massive MIMO	PS	Sum rate (Dai et al., 2019): the maximum achievable sum rate of system
CR	Nonlinear EH model	Secrecy performance (Zhou FH et al., 2018a): security of the primary network and the minimum transmitted power

7.4 Imperfect SIC modeling

SIC is a key component at the receiver end to mitigate co-channel interference caused by nonorthogonal transmission. In 5G and B5G wireless networks, most nodes are IoT devices with limited capability. As a result, decoding error may occur during the processing of interference cancellation. Especially, since short packets are used in IoT communications, there exists a high probability of decoding error. Due to error propagation in imperfect SIC, the residual interference is strong. Fortunately, it is possible to alleviate the impact of imperfect SIC by multiple-antenna techniques, e.g., spatial beamforming, power allocation, and user clustering. In order to effectively alleviate the impact of imperfect SIC, it is imperative to build an accurate model of imperfect SIC. Currently, a linear model is adopted in multiple-antenna NOMA systems. However, the linear model might be not sufficiently accurate. Therefore, we should accurately study imperfect SIC modeling for multiple-antenna NOMA systems.

7.5 Cell-free NOMA

5G and B5G wireless networks are required to provide a wide coverage for a massive number of UEs. However, the current centralized cell architecture leads to a severe far-near effect. Specifically, the central UEs receive strong signals and weak interference, while edge UEs receive weak signals and strong interference. As a result, the quality of the received signals at edge UEs cannot satisfy the requirements of QoS, and then the cell coverage is limited. To solve this problem, a feasible way is the use of cell-free architecture. In particular, the BS antennas are distributed over the whole area, and they are connected to a central processing unit. Hence, the access distance can be shortened, and the coverage is wide. However, the design of cell-free NOMA is still an open issue.

8 Conclusions

This study has provided a review on multiple-antenna techniques in NOMA systems from theory to technique. First, we presented an investigation of various key techniques of multiple-antenna NOMA systems, including CSI acquisition, user clustering, superposition coding, and SIC. Then, we reviewed the issues related to multiple-antenna techniques for two-user NOMA systems. Subsequently, we presented an overview on multiple-antenna NOMA transmission in multiuser and massive-connectivity scenarios. Then, we went ahead to provide a review of multiple-antenna NOMA in various heterogeneous networks. Finally, we discussed the potential challenges for NOMA based on multiple-antenna techniques and pointed out some possible future research directions.

Compliance with ethics guidelines

Fei-yan TIAN and Xiao-ming CHEN declare that they have no conflict of interest.

References

Alavi F, Cumanan K, Ding ZG, et al., 2017. Robust beamforming techniques for non-orthogonal multiple access systems with bounded channel uncertainties. *IEEE Commun Lett*, 21(9):2033-2036. https://doi.org/10.1109/LCOMM.2017.2702580

Ali E, Ismail M, Nordin R, et al., 2017. Beamforming techniques for massive MIMO systems in 5G: overview, classification, and trends for future research. *Front Inform Technol Electron Eng*, 18(6):753-772. https://doi.org/10.1631/FITEE.1601817

Ali MS, Tabassum H, Hossain E, 2016a. Dynamic user clustering and power allocation for uplink and downlink non-orthogonal multiple access (NOMA) systems. *IEEE Access*, 4:6325-6343. https://doi.org/10.1109/ACCESS.2016.2604821

Ali MS, Hossain E, Kim DI, 2016b. Non-orthogonal multiple access (NOMA) for downlink multiuser MIMO systems: user clustering, beamforming, and power allocation. *IEEE Access*, 5:565-577. https://doi.org/10.1109/ACCESS.2016.2646183

Amin SH, Mehana AH, Soliman SS, et al., 2018. Power allocation for maximum MIMO-NOMA system user-rate. Proc Globecom Workshops, p.1-6. https://doi.org/10.1109/GLOCOMW.2018.8644505

Arzykulov S, Tsiftsis TA, Nauryzbayev G, et al., 2019. Outage performance of cooperative underlay CR-NOMA with imperfect CSI. *IEEE Commun Lett*, 23(1):176-179. https://doi.org/10.1109/LCOMM.2018.2878730

Bai L, Zhu L, Yu Q, et al., 2019. Transmit power minimization for vector-pertubation based NOMA systems: a sub-optimal beamforming approach. *IEEE Trans Wirel Commun*, 18(5):2679-2692. https://doi.org/110.1109/TWC.2019.2906909

Cai W, Lv G, Jin Y, 2017. Half-ZF beamforming scheme for downlink two-user multiple input single output-based non-orthogonal multiple access systems. *IET Commun*, 11(10):1633-1640. https://doi.org/10.1049/iet-com.2017.0018

Catarinucci L, de Donno D, Mainetti L, 2015. An IoT-aware architecture for smart healthcare systems. *IEEE Int Things J*, 2(6):515-526. https://doi.org/10.1109/JIOT.2015.2417684

Celik A, Al-Qahtani FS, Radaydeh RM, et al., 2017. Cluster formation and joint power-bandwidth allocation for imperfect NOMA in DL-HetNets. Proc Global Communications Conf, p.1-6. https://doi.org/10.1109/GLOCOM.2017.8254637

Celik A, Tsai MC, Radaydeh RM, et al., 2019. Distributed cluster formation and power-bandwidth allocation for imperfect NOMA in DL-HetNets. *IEEE Trans Commun*, 67(2):1677-1692. https://doi.org/10.1109/TCOMM.2018.2879508

Chen C, Cai WB, Cheng X, et al., 2017. Low complexity beamforming and user selection schemes for 5G MIMO-NOMA systems. *IEEE J Sel Areas Commun*, 35(12):2708-2722. https://doi.org/10.1109/JSAC.2017.2727229

Chen J, Chen XM, Gerstacker WH, et al., 2016. Resource allocation for a massive MIMO relay aided secure communication. *IEEE Trans Inform Forens Secur*, 11(8):1700-1711. https://doi.org/10.1109/TIFS.2016.2551685

Chen SZ, Kang SL, 2018. A tutorial on 5G and the progress in China. *Front Inform Technol Electron Eng*, 19(3):309-321. https://doi.org/10.1631/FITEE.1800070

Chen X, 2019. Massive Access for Cellular Internet of Things: Theory and Technique. Springer Press, Germany.

Chen X, Chen HH, 2014. Physical layer security in multi-cell MISO downlink with incomplete CSI—a unified secrecy performance analysis. *IEEE Trans Signal Process*, 62(23):6286-6297. https://doi.org/10.1109/TSP.2014.2362890

Chen X, Yuen C, 2014. Performance analysis and optimization for interference alignment over MIMO interference channels with limited feedback. *IEEE Trans Signal Process*, 62(7):1785-1795. https://doi.org/10.1109/TSP.2014.2304926

Chen X, Zhang Y, 2017. Mode selection in MU-MIMO downlink networks: a physical layer security perspective. *IEEE Syst J*, 11(2):1128-1136. https://doi.org/10.1109/JSYST.2015.2413843

Chen X, Gong FK, Li G, et al., 2018. User pairing and pairing scheduling in massive MIMO-NOMA systems. *IEEE Commun Lett*, 22(4):788-791. https://doi.org/10.1109/LCOMM.2017.2776206

Chen XM, Chen HH, 2013. Interference-aware resource control in multi-antenna cognitive ad hoc networks with heterogeneous delay constraints. *IEEE Commun Lett*, 17(6):1184-1187. https://doi.org/10.1109/LCOMM.2013.042313.130481

Chen XM, Jia RD, 2018. Exploiting rateless coding for massive access. *IEEE Trans Veh Technol*, 67(11):11253-11257. https://doi.org/10.1109/TVT.2018.2866279

Chen XM, Yuen C, 2013. Efficient resource allocation in rateless coded MU-MIMO cognitive radio network with QoS provisioning and limited feedback. *IEEE Trans Veh Technol*, 62(1):395-399. https://doi.org/10.1109/TVT.2012.2219568

Chen XM, Zhang ZY, 2010. Exploiting channel angular domain information for precoder design in distributed antenna system. *IEEE Trans Signal Process*, 58(11):5791-5801. https://doi.org/10.1109/TSP.2010.2062508

Chen XM, Zhang ZY, Chen HH, 2010. On distributed antenna system with limited feedback precoding-opportunities and challenges. *IEEE Wirel Commun*, 17(2):80-88. https://doi.org/10.1109/MWC.2010.5450664

Chen XM, Zhang ZY, Chen SL, et al., 2012. Adaptive mode selection for multiuser MIMO downlink employing rateless codes with QoS provisioning. *IEEE Trans Wirel Commun*, 11(2):790-799. https://doi.org/10.1109/TWC.2011.120511.110748

Chen XM, Wang XM, Chen XF, 2013. Energy-efficient optimization for wireless information and power transfer in large-scale MIMO systems employing energy beamforming. *IEEE Wirel Commun Lett*, 2(6):667-670. https://doi.org/10.1109/WCL.2013.092813.130514

Chen XM, Chen HH, Meng WX, 2014a. Cooperative communications for cognitive radio networks—from theory to applications. *IEEE Commun Surv Tutor*, 16(3):1180-1193. https://doi.org/10.1109/SURV.2014.021414.00066

Chen XM, Yuen C, Zhang ZY, 2014b. Wireless energy and information transfer tradeoff for limited feedback multi-antenna systems with energy beamforming. *IEEE Trans Veh Technol*, 63(1):407-412. https://doi.org/10.1109/TVT.2013.2274800

Chen XM, Zhang ZY, Chen HH, et al., 2015a. Enhancing wireless information and power transfer by exploiting multi-antenna techniques. *IEEE Commun Mag*, 53(4):133-141. https://doi.org/10.1109/MCOM.2015.7081086

Chen XM, Lei L, Zhang HZ, et al., 2015b. Large-scale MIMO relaying techniques for physical layer security: AF or DF? *IEEE Trans Wirel Commun*, 14(9):5135-5146. https://doi.org/10.1109/TWC.2015.2433291

Chen XM, Zhong CJ, Yuen C, et al., 2015c. Multi-antenna relay aided wireless physical layer security. *IEEE Commun Mag*, 53(12):40-46. https://doi.org/10.1109/MCOM.2015.7355564

Chen XM, Ng DWK, Chen HH, 2016. Secrecy wireless information and power transfer: challenges and opportunities. *IEEE Wirel Commun*, 23(2):54-61. https://doi.org/10.1109/MWC.2016.7462485

Chen XM, Zhang ZY, Zhong CJ, et al., 2017a. Exploiting multiple-antenna techniques for non-orthogonal multiple access. *IEEE J Sel Areas Commun*, 35(10):2207-2220. https://doi.org/10.1109/JSAC.2017.2724420

Chen XM, Ng DWK, Gerstacker W, et al., 2017b. A survey on multiple-antenna techniques for physical layer security. *IEEE Commun Surv Tutor*, 19(2):1027-1053. https://doi.org/10.1109/COMST.2016.2633387

Chen XM, Jia RD, Ng DWK, 2018a. The application of relay to massive non-orthogonal multiple access. *IEEE Trans Commun*, 66(11):5168-5180. https://doi.org/10.1109/TCOMM.2018.2854578

Chen XM, Zhang ZY, Zhong CJ, et al., 2018b. Exploiting inter-user interference for secure massive non-orthogonal multiple access. *IEEE J Sel Areas Commun*, 36(4):788-801. https://doi.org/10.1109/JSAC.2018.2825058

Chen XM, Zhang ZY, Zhong CJ, et al., 2018c. Fully non-orthogonal communication for massive access. *IEEE Trans Commun*, 66(4):1717-1731. https://doi.org/10.1109/TCOMM.2017.2779150

Chen XM, Jia R, Ng DWK, 2019. On the design of massive non-orthogonal multiple access with imperfect successive interference cancellation. *IEEE Trans Commun*, 67(3):2539-2551. https://doi.org/10.1109/TCOMM.2018.2884476

Chen ZL, Sohrabi F, Yu W, 2018. Sparse activity detection for massive connectivity. *IEEE Trans Signal Process*, 66(7):1890-1904. https://doi.org/10.1109/TSP.2018.2795540

Chen ZY, Ding ZG, Dai XC, 2016a. Beamforming for combating inter-cluster and intra-cluster interfernece in hybrid NOMA systems. *IEEE Access*, 4:4452-4463. https://doi.org/10.1109/ACCESS.2016.2598380

Chen ZY, Ding ZG, Xu P, et al., 2016b. Optimal precoding for a QoS optimization problem in two-user MISO-NOMA downlink. *IEEE Commun Lett*, 20(6):1263-1266. https://doi.org/10.1109/LCOMM.2016.2555907

Cheng HV, Bjornson E, Larsson EG, 2018. Performance analysis of NOMA in training-based multiuser MIMO systems. *IEEE Trans Commun*, 17(1):372-385. https://doi.org/10.1109/TWC.2017.2767030

Chitti K, Rusek F, Tumula C, 2017. Bandwidth minimization under probabilistic constraints and statistical CSI for NOMA. Proc 86th Vehicular Technology Conf, p.1-5. https://doi.org/10.1109/VTCFall.2017.8288401

Choi J, 2015. Minimum power multicast beamforming with superposition coding for multiresolution broadcast and application to NOMA systems. *IEEE Trans Commun*, 63(3):791-800. https://doi.org/10.1109/TCOMM.2015.2394393

Choi J, 2016a. Power allocation for max-sum rate and max-min rate proportional fairness in NOMA. *IEEE Commun Lett*, 20(10):2055-2058. https://doi.org/10.1109/LCOMM.2016.2596760

Choi J, 2016b. On the power allocation for MIMO-NOMA systems with layered transmission. *IEEE Trans Wirel Commun*, 15(5):3226-3237. https://doi.org/10.1109/TWC.2016.2518182

Chraiti M, Ghrayeb A, Assi C, 2018. A NOMA scheme for a two-user MISO downlink channel with unknow CSIT. *IEEE Trans Wirel Commun*, 17(10):6775-6789. https://doi.org/10.1109/TWC.2018.2864215

Cui JJ, Liu YW, Ding ZG, et al., 2018a. Optimal user scheduling and power allocation for millimeter wave NOMA systems. *IEEE Trans Wirel Commun*, 17(3):1502-1517. https://doi.org/10.1109/TWC.2017.2779504

Cui JJ, Ding ZG, Fan PZ, et al., 2018b. Unsupervised machine learning based user clustering in mmwave-NOMA systems. *IEEE Trans Wirel Commun*, 17(11):7425-7440. https://doi.org/10.1109/TWC.2018.2867180

Cui JJ, Ding ZG, Fan PZ, 2018c. Outage probability constrained MIMO-NOMA design under imperfect CSI. *IEEE Trans Wirel Commun*, 17(12):8239-8255. https://doi.org/10.1109/TWC.2018.2875490

Dai JL, Sun L, Yang CY, 2017. On the average rate and power allocation of uplink multi-antenna NOMA systems. Proc 86th Vehicular Technology Conf, p.1-5. https://doi.org/10.1109/VTCFall.2017.8288076

Dai LL, Wang BC, Yuan YF, 2015. Non-orthogonal multiple access for 5G: solutions, challenges, opportunities, and future research trends. *IEEE Commun Mag*, 53(9):74-81. https://doi.org/10.1109/MCOM.2015.7263349

Dai LL, Wang BC, Ding ZG, et al., 2018. A survey of non-orthogonal multiple access for 5G. *IEEE Commun Surv Tutor*, 20(3):2294-2323. https://doi.org/10.1109/COMST.2018.2835558

Dai LL, Wang BC, Peng MG, et al., 2019. Hybrid precoding-based millimeter-wave massive MIMO-NOMA with simultaneous wireless information and power transfer. *IEEE J Sel Areas Commun*, 37(1):131-141. https://doi.org/10.1109/JSAC.2018.2872364

Dai W, Liu YJ, Rider B, 2008. Quantization bounds on Grassmann manifolds and applications to MIMO communications. *IEEE Trans Inform Theory*, 54(3):1108-1123. https://doi.org/10.1109/TIT.2007.915691

Ding JF, Cai J, Yi CY, 2019. An improved coalition game approach for MIMO-NOMA clustering integrating beamforming and power allocation. *IEEE Trans Veh Technol*, 68(2):1672-1687. https://doi.org/10.1109/TVT.2018.2889694

Ding ZG, Yang Z, Fan PZ, et al., 2014. On the performance of non-orthogonal multiple access in 5G systems with randomly deployed users. *IEEE Signal Process Lett*, 21(12):1501-1505. https://doi.org/10.1109/LSP.2014.2343971

Ding ZG, Adachi F, Poor HV, 2016a. The application of MIMO to non-orthogonal multiple access. *IEEE Trans Wirel Commun*, 15(1):537-552. https://doi.org/10.1109/TWC.2015.2475746

Ding ZG, Schober R, Poor HV, 2016b. A general MIMO framework for NOMA downlink and uplink transmission based on signal alignment. *IEEE Trans Wirel Commun*, 15(6):4438-4454. https://doi.org/10.1109/TWC.2016.2542066

Ding ZG, Fan PZ, Poor HV, 2016c. Impact of user pairing on 5G nonorthogonal multiple-access downlink transmissions. *IEEE Trans Veh Technol*, 65(8):6010-6023. https://doi.org/10.1109/TVT.2015.2480766

Ding ZG, Dai LL, Schober R, et al., 2017a. NOMA meets finite resolution analog beamforming in massive MIMO and milimeter-wave networks. *IEEE Commun Lett*, 21(8):1879-1882. https://doi.org/10.1109/LCOMM.2017.2700846

Ding ZG, Fan PZ, Poor HV, 2017b. Random beamforming in millimeter-wave NOMA networks. *IEEE Access*, 5:7667-7681. https://doi.org/10.1109/ACCESS.2017.2673248

Ding ZG, Zhao ZY, Peng MG, et al., 2017c. On the spectral efficiency and security enhancements of NOMA assisted multicast-unicast streaming. *IEEE Trans Commun*, 65(7):3151-3163. https://doi.org/10.1109/TCOMM.2017.2696527

Ding ZG, Lei XF, Karagiannidis GK, 2017d. A survey on non-orthogonal multiple access for 5G networks: research challenges and future trends. *IEEE J Sel Areas Commun*, 35(10):2181-2195. https://doi.org/10.1109/JSAC.2017.2725519

Ding ZG, Xu M, Chen Y, et al., 2018. Embracing non-orthogonal multiple access in future wireless networks. *Front Inform Technol Electron Eng*, 19(3):322-339. https://doi.org/10.1631/FITEE.1800051

Do TN, da Costa DB, Duong TQ, et al., 2018. Improving the performance of cell-edge users in MISO-NOMA systems using TAS and SWIPT-based cooperative transmissions. *IEEE Trans Green Commun Netw*, 2(1):49-62. https://doi.org/10.1109/TGCN.2017.2777510

Du Y, Cheng C, Dong BH, et al., 2018a. Block-sparsity-based multiuser detection for uplink grant-free NOMA. *IEEE Trans Wirel Commun*, 17(12):7894-7909. https://doi.org/10.1109/TWC.2018.2872594

Du Y, Dong BH, Zhu WY, et al., 2018b. Joint channel estimation and multiuser detection for uplink grant-free NOMA. *IEEE Wirel Commun Lett*, 7(4):682-685. https://doi.org/10.1109/LWC.2018.2810278

Elijah O, Leow CY, Rahman TA, et al., 2016. A comprehensive survey of pilot contamination in massive MIMO-5G system. *IEEE Commun Surv Tutor*, 18(2):905-923. https://doi.org/10.1109/COMST.2015.2504379

Fang F, Zhang HJ, Cheng JL, et al., 2017. Joint user scheduling and power allocation optimization for energy-efficient NOMA systems with imperfect CSI. *IEEE J Sel Areas Commun*, 35(12):2874-2885. https://doi.org/10.1109/JSAC.2017.2777672

Gomez G, Martin-Vega FJ, Lopez-Martinez FJ, et al., 2019. Physical layer security in uplink NOMA multi-antenna systems with randomly distributed eavesdroppers. *IEEE Access*, 7:70422-70435. https://doi.org/10.1109/ACCESS.2019.2920578

Hanif MF, Ding ZG, Ratnarajah T, et al., 2016. A minorization-maximization method for optimizing sum rate in the downlink of non-orthogonal multiple access systems. *IEEE Trans Signal Process*, 64(1):76-88. https://doi.org/10.1109/TSP.2015.2480042

Haupt RL, Rahmat-Samii Y, 2015. Antenna array developments: a perspective on the past, present and future. *IEEE Anten Propag Mag*, 57(1):86-96. https://doi.org/10.1109/MAP.2015.2397154

He B, Liu A, Yang N, et al., 2017. On the design of secure non-orthogonal multiple access systems. *IEEE J Sel Areas Commun*, 35(10):2196-2206. https://doi.org/10.1109/JSAC.2017.2725698

Hosseini K, Yu W, Adve RS, 2014. Large-scale MIMO versus network MIMO for multicell interference mitigation. *IEEE J Sel Topics Signal Process*, 8(5):930-941. https://doi.org/10.1109/JSTSP.2014.2327594

Hoydis J, ten Brink S, Debbah M, 2013. Massive MIMO in the UL/DL of cellular networks: how many antennas do we need? *IEEE J Sel Areas Commun*, 31(2):160-171. https://doi.org/10.1109/JSAC.2013.130205

Hu CY, Wang YS, Hong YWP, et al., 2017. MMSE hybrid beamforming for weighted sum rate maximization in NOMA systems. Proc Global Communications Conf, p.1-6. https://doi.org/10.1109/GLOCOM.2017.8254539

Hu XL, Zhong CJ, Han Y, et al., 2019. Angle-domain mmWave MIMO NOMA systems: analysis and design. Proc Int Conf on Communications, p.1-6. https://doi.org/10.1109/ICC.2019.8761180

Islam SMR, Avazov N, Dobre OA, 2017. Power-domain non-orthogonal multiple access (NOMA) in 5G systems: potentials and challenges. *IEEE Commun Surv Tutor*, 19(2):721-742. https://doi.org/10.1109/COMST.2016.2621116

Jia RD, Chen XM, Zhong CJ, et al., 2019. Design of non-orthogonal beamspace multiple access for cellular Internet-of-Things. *IEEE J Sel Top Signal Process*, 13(3):538-552. https://doi.org/10.1109/JSTSP.2019.2898331

Kader F, Shin SY, 2016. Cooperative spectrum sharing with space time block coding and non-orthogonal multiple access. Proc 8th Int Conf on Ubiquitous and Future Networks, p.490-494. https://doi.org/10.1109/ICUFN.2016.7537080

Kang JM, Kim IM, 2018. Optimal user grouping for downlink NOMA. *IEEE Wirel Commun Lett*, 7(5):724-727. https://doi.org/10.1109/LWC.2018.2815683

Khisti A, Wornell GW, 2010. Secure transmission with multiple antennas I: the MISOME wiretap channel. *IEEE Trans Inform Theory*, 56(7):3088-3014. https://doi.org/10.1109/TIT.2010.2048445

Kim B, Lim S, Kim H, et al., 2013. Non-orthogonal multiple access in a downlink multiuser beamforming system. Proc Military Communications Conf, p.1278-1283. https://doi.org/10.1109/MILCOM.2013.218

Kim J, Koh J, Kang J, et al., 2015. Design of user clustering and precoding for downlink non-orthogonal multiple access (NOMA). Proc Military Communications Conf, p.1170-1175. https://doi.org/10.1109/MILCOM.2015.7357604

Lei HJ, Zhang JM, Park KH, et al., 2018. Secrecy outage of max-min TAS scheme in MIMO-NOMA systems. *IEEE Trans Veh Technol*, 67(8):6981-6990. https://doi.org/10.1109/TVT.2018.2824310

Lei L, Yuan D, Ho CK, et al., 2016. Power and channel allocation for non-orthogonal multiple access in 5G systems: tractability and computation. *IEEE Trans Wirel Commun*, 15(12):8580-8594. https://doi.org/10.1109/TWC.2016.2616310

Li F, Zhang QT, 2007. Transmission strategy for MIMO correlated Rayleigh fading channels with mutual coupling. Proc Int Conf on Communications, p.1030-1035. https://doi.org/10.1109/ICC.2007.175

Li YQ, Jiang M, Zhang Q, et al., 2017. Secure beamforming in downlink MISO nonorthogonal multiple access systems. *IEEE Trans Veh Technol*, 66(8):7563-7567. https://doi.org/10.1109/TVT.2017.2658563

Li YQ, Jiang M, Zhang Q, et al., 2018. Cooperative non-orthogonal multiple access in multiple-input-multiple-output channels. *IEEE Trans Wirel Commun*, 17(3):2068-2079. https://doi.org/10.1109/TWC.2017.2788413

Liang W, Ding ZG, Li YH, et al., 2017. User pairing for downlink non-orthogonal multiple access networks using matching algorithm. *IEEE Trans Commun*, 65(12):5319-5332. https://doi.org/10.1109/TCOMM.2017.2744640

Liu F, Petrova M, 2018. Dynamic power allocation for downlink multi-carrier NOMA systems. *IEEE Commun Lett*, 22(9):1930-1933. https://doi.org/10.1109/LCOMM.2018.2852655

Liu JX, Xiong K, Lu Y, et al., 2008. SWIPT-enabled NOMA networks with full-duplex relaying. Proc Global Communications Conf, p.1-6. https://doi.org/10.1109/GLOCOM.2018.8647987

Liu L, Larsson EG, Yu W, et al., 2018. Sparse signal processing for grant-free massive connectivity: a future paradigm for random access protocols in the Internet of Things. *IEEE Signal Process Mag*, 35(5):88-99. https://doi.org/10.1109/MSP.2018.2844952

Liu L, Chi YH, Yuen C, et al., 2019. Capacity-achieving MIMO-NOMA: iterative LMMSE detection. *IEEE Trans Signal Process*, 67(7):1758-1773. https://doi.org/10.1109/TSP.2019.2896242

Liu PL, Li Y, Cheng W, et al., 2019. Energy-efficient power allocation for millimeter wave beamspace MIMO-NOMA systems. *IEEE Access*, 7:114582-114592. https://doi.org/10.1109/ACCESS.2019.2935495

Liu X, Liu YN, Wang XB, et al., 2017. Highly efficient 3-D resource allocation techniques in 5G for NOMA-enabled massive MIMO and relaying systems. *IEEE J Sel Areas Commun*, 35(12):2785-2797. https://doi.org/10.1109/JSAC.2017.2726378

Liu YW, Elkashlan M, Ding ZG, et al., 2016. Fairness of user clustering in MIMO non-orthogonal multiple access systems. *IEEE Commun Lett*, 20(7):1465-1468. https://doi.org/10.1109/LCOMM.2016.2559459

Liu YW, Qin ZJ, Elkashlan M, et al., 2017. Nonorthogonal multiple access for 5G and beyond. *Proc IEEE*, 105(12):2347-2381. https://doi.org/10.1109/JPROC.2017.2768666

Liu ZX, Lei L, Zhang NB, et al., 2017. Joint beamforming and power optimization with iterative user clustering for MISO-NOMA systems. *IEEE Access*, 5:6872-6884. https://doi.org/10.1109/ACCESS.2017.2700018

Love DJ, Heath RW, Lau VKN, et al., 2008. An overview of limited feedback in wireless communication systems. *IEEE J Sel Areas Commun*, 26(8):1341-1365. https://doi.org/10.1109/JSAC.2008.081002

Lu L, Li GY, Swindlehurst AL, et al., 2014. An overview of massive MIMO: benefits and challenges. *IEEE J Sel Top Signal Process*, 8(5):742-758. https://doi.org/10.1109/JSTSP.2014.2317671

Lv L, Chen J, Ni Q, 2016. Cooperative non-orthogonal multiple access in cognitive radio. *IEEE Commun Lett*, 20(10):2059-2062. https://doi.org/10.1109/LCOMM.2016.2596763

Lv L, Ni Q, Ding ZG, et al., 2017. Application of non-orthogonal multiple access in cooperative spectrum-sharing networks over Nakagami-m fading channels. *IEEE Trans Veh Technol*, 66(6):5506-5511. https://doi.org/10.1109/TVT.2016.2627559

Lv L, Ding ZG, Ni Q, et al., 2018a. Secure MISO-NOMA transmission with artificial noise. *IEEE Trans Veh Technol*, 67(7):6700-6705. https://doi.org/10.1109/TVT.2018.2811733

Lv L, Yang L, Jiang H, et al., 2018b. When NOMA meets multiuser cognitive radio: opportunistic cooperation and user scheduling. *IEEE Trans Veh Technol*, 67(7):6679-6684. https://doi.org/10.1109/TVT.2018.2805638

Lv T, Ma Y, Zeng J, et al., 2018. Millimeter-wave NOMA transmission in cellular M2M communications for Internet of Things. *IEEE Internet Things J*, 5(3):1989-2000. https://doi.org/10.1109/JIOT.2018.2819645

Manglayev T, Kizilirmak RC, Kho YH, et al., 2017. NOMA with imperfect SIC implementation. Proc 17th Int Conf on Smart Technologies, p.22-25. https://doi.org/10.1109/EUROCON.2017.8011071

Mei XD, Wu KL, 2018. How low does mutual coupling need to be for MIMO antennas. Proc Int Symp on Antennas and Propagation & USNC/URSI National Radio Science Meeting, p.1579-1580. https://doi.org/10.1109/APUSNCURSINRSM.2018.8608539

Mi D, Dianati M, Zhang L, et al., 2017. Massive MIMO performance with imperfect channel reciprocity and channel estimation error. *IEEE Trans Wirel Commun*, 65(9):3734-3749. https://doi.org/10.1109/TCOMM.2017.2676088

Mitra R, Bhatia V, 2017. Precoded Chebyshev-NLMS-based pre-distorter for nonlinear LED compensation in NONMA-VLC. *IEEE Trans Commun*, 65(11):4845-4856. https://doi.org/10.1109/TCOMM.2017.2736548

Moltafet M, Yamchi NM, Javan MR, et al., 2018a. Comparison study between PD-NOMA and SCMA. *IEEE Trans Veh Technol*, 67(2):1830-1834. https://doi.org/10.1109/TVT.2017.2759910

Nguyen NP, Dobre OA, Nguyen LD, et al., 2019. Secure downlink massive MIMO NOMA network in the presence of a multiple-antenna eavesdropper. Proc Int Conf on Communications, p.1-6. https://doi.org/10.1109/ICC.2019.8761708

Nguyen TS, Duy HHK, Nguyen H, et al., 2018. Throughput analysis in relaying cooperative systems considering time-switching with NOMA. Proc 41st Int Conf on Telecommunications and Signal Processing, p.1-4. https://doi.org/10.1109/TSP.2018.8441516

Nguyen VD, Tuan HD, Duong TQ, et al., 2017a. Joint fractional time allocation and beamforming for downlink multiuser MISO systems. *IEEE Commun Lett*, 21(12):2650-2653. https://doi.org/10.1109/LCOMM.2017.2747544

Nguyen VD, Tuan HD, Duong TQ, et al., 2017b. Precoder design for signal superposition in MIMO-NOMA multicell networks. *IEEE J Sel Areas Commun*, 35(12):2681-2695. https://doi.org/10.1109/JSAC.2017.2726007

Nomikos N, Michailidis ET, Trakadas P, et al., 2019. Flex-NOMA: exploiting buffer-aided relay selection for massive connectivity in the 5G uplink. *IEEE Access*, 7:88743-88755. https://doi.org/10.1109/ACCESS.2019.2926770

Palattella MR, Dohler M, Grieco A, 2016. Internet of Things in the 5G era: enablers, architecture, and business models. *IEEE J Sel Areas Commun*, 34(3):510-527. https://doi.org/10.1109/JSAC.2016.2525418

Peng JJ, Chen W, Ai B, et al., 2017. Joint optimization of constellation with mapping matrix for SCMA codebook design. *IEEE Signal Process Lett*, 24(3):264-268. https://doi.org/10.1109/LSP.2017.2653845

Qi Q, Chen XM, 2019. Wireless powered massive access for cellular Internet of Things with imperfect SIC and nonlinear EH. *IEEE Int Things J*, 6(2):3110-3120. https://doi.org/10.1109/JIOT.2018.2878860

Sedaghat MA, Müller RR, 2018. On user pairing in uplink NOMA. *IEEE Trans Wirel Commun*, 17(5):3474-3486. https://doi.org/10.1109/TWC.2018.2815005

Senel K, Larsson EG, 2018. Grant-free massive MTC-enabled massive MIMO: a compressive sensing approach. *IEEE Trans Commun*, 66(12):6164-6175. https://doi.org/10.1109/TCOMM.2018.2866559

Seo J, Sung Y, 2018. Beam design and user scheduling for nonorthogonal multiple access with multiple antenna based on Pareto optimality. *IEEE Trans Signal Process*, 66(11):2876-2891. https://doi.org/10.1109/TSP.2018.2821635

Shao XD, Chen XM, Zhong CJ, et al., 2019. A unified design of massive access for cellular Internet of Things. *IEEE Int Things J*, 6(2):3934-3947. https://doi.org/10.1109/JIOT.2019.2893376

Shi Z, Yang GH, Fu YR, et al., 2018. Performance analysis of MIMO-NOMA systems with randomly deployed users. Proc Global Communications Conf, p.1-7. https://doi.org/10.1109/GLOCOM.2018.8647702

Shin W, Vaezi M, Lee B, et al., 2017a. Coordinated beamforming for multi-cell MIMO-NOMA. *IEEE Commun Lett*, 21(1):84-87. https://doi.org/10.1109/LCOMM.2016.2615097

Shin W, Vaezi M, Lee B, et al., 2017b. Non-orthogonal multiple access in multi-cell networks: theory, performance, and practical challenges. *IEEE Commun Mag*, 55(10):176-183. https://doi.org/10.1109/MCOM.2017.1601065

Shin W, Yang H, Vaezi M, et al., 2017c. Relay-aided NOMA in uplink cellular networks. *IEEE Signal Process Lett*, 24(12):1842-1846. https://doi.org/10.1109/LSP.2017.2755049

Sohrabi F, Yu W, 2016. Hybrid digital and analog beamforming design for large-scale antenna arrays. *IEEE J Sel Top Signal Process*, 10(3):501-513. https://doi.org/10.1109/JSTSP.2016.2520912

Sun Q, Han SF, Chin-Lin I, et al., 2015. On the ergodic capacity of MIMO NOMA systems. *IEEE Wirel Commun Lett*, 4(4):405-408. https://doi.org/10.1109/LWC.2015.2426709

Sun YS, Ding ZG, Dai XC, et al., 2018. A feasibility study on network NOMA. *IEEE Trans Commun*, 66(9):4303-4317. https://doi.org/10.1109/TCOMM.2018.2825420

Tian MX, Zhang Q, Zhao S, et al., 2018. Robust beamforming in downlink MIMO NOMA networks using cutting-set method. *IEEE Commun Lett*, 22(3):574-577. https://doi.org/10.1109/LCOMM.2017.2773088

Timotheou S, Krikidis I, 2015. Fairness for non-orthogonal multiple access in 5G systems. *IEEE Signal Process Lett*, 22(10):1647-1651. https://doi.org/10.1109/LSP.2015.2417119

Tong D, Ding YH, Liu Y, et al., 2019. A MIMO-NOMA framework with complex-valued power coefficients. *IEEE Trans Veh Technol*, 68(3):2244-2259. https://doi.org/10.1109/TVT.2018.2890546

TR G, 2015. Technical Specification Group GSM/EDGE Radio Access Network; Cellular System Support for Ultra-low Complexity and Low Throughput Internet of Things (CIoT) TR 45.820. 3rd Generation Partnership Project, 3GPP.

Tsai YR, Wei HA, 2018. Quality-balanced user clustering schemes for non-orthogonal multiple access systems. *IEEE Commun Lett*, 22(1):113-116. https://doi.org/10.1109/LCOMM.2017.2766618

Vaezi M, Schober R, Ding ZG, et al., 2019. Non-orthogonal multiple access: common myths and critical questions. *IEEE Wirel Commun*, 26(5):174-180. https://doi.org/10.1109/MWC.2019.1800598

Varshney LR, 2008. Transporting information and energy simultaneously. Proc Int Symp on Information Theory, p.1612-1616. https://doi.org/10.1109/ISIT.2008.4595260

Wan D, Chen D, Song B, et al., 2018. From IoT to 5G I-IoT: te next generation IoT-based intelligent algorithms and 5G technologies. *IEEE Commun Mag*, 56(10):114-120. https://doi.org/10.1109/MCOM.2018.1701310

Wan DH, Wen MW, Ji F, et al., 2018. Cooperative NOMA systems with partial channel state information over Nakagami-m fading channels. *IEEE Trans Commun*, 66(3):947-958. https://doi.org/10.1109/TCOMM.2017.2772273 ‘

Wang BC, Dai LL, Zhang Y, et al., 2016. Dynamic compressive sensing-based multi-user detection for uplink grant-free NOMA. *IEEE Commun Lett*, 20(11):2320-2323. https://doi.org/10.1109/LCOMM.2016.2602264

Wang BC, Dai LL, Wang ZC, et al., 2017. Spectrum and energy-efficient beamspace MIMO-NOMA for millimeter-wave communications using lens antenna array. *IEEE J Sel Areas Commun*, 35(10):2370-2382. https://doi.org/10.1109/JSAC.2017.2725878

Wang CS, Wang Y, Wang W, et al., 2017. Electromechanical coupling based influence of structural error on radiation and scattering performance of array antennas. *Electron Lett*, 53(14):904-906. https://doi.org/10.1049/el.2017.1072

Wang CS, Wang Y, Zhou JZ, et al., 2018. Compensation method for distorted planar array antennas based on structural-electromagnetic coupling and fast Fourier transform. *IET Microw Anten Propag*, 12(6):954-962. https://doi.org/10.1049/iet-map.2017.0814

Wang H, Zhang ZY, Chen XM, 2017. Resource allocation for downlink joint space-time and power domain non-orthogonal multiple access. Proc 9th Int Conf on Wireless Communications and Signal Processing, p.1-6. https://doi.org/10.1109/WCSP.2017.8171094

Wang H, Zhang RB, Song RF, et al., 2018. A novel power minimization precoding scheme for MIMO-NOMA uplink systems. *IEEE Commun Lett*, 22(5):1106-1109. https://doi.org/10.1109/LCOMM.2018.2812786

Wang JH, Peng Q, Huang YM, et al., 2017. Convexity of weighted sum rate maximization in NOMA systems. *IEEE Signal Process Lett*, 24(9):1323-1327. https://doi.org/10.1109/LSP.2017.2722546

Wang XS, Wang JT, He LZ, et al., 2018. Outage analysis for downlink NOMA with statistical channel state information. *IEEE Wirel Commun Lett*, 7(2):142-145. https://doi.org/10.1109/LWC.2017.2761343

Wang XY, Jia M, Guo Q, et al., 2019. Full-duplex relaying cognitive radio network with cooperative nonorthogonal multiple access. *IEEE Syst J*, 13(4):3897-3908. https://doi.org/10.1109/JSYST.2019.2927509

Wei C, Liu HP, Zhang ZC, et al., 2017. Approximate message passing-based joint user activity and data detection for NOMA. *IEEE Commun Lett*, 21(3):640-643. https://doi.org/10.1109/LCOMM.2016.2624297

Wei F, Chen W, 2017. Low complexity iterative receiver design for sparse code multiple access. *IEEE Trans Commun*, 65(2):621-634. https://doi.org/10.1109/TCOMM.2016.2631468

Wei ZQ, Ng DWK, Yuan JH, 2016a. Power-efficient resource allocation for MC-NOMA with statistical channel state information. Proc Global Communications Conf, p.1-7. https://doi.org/10.1109/GLOCOM.2016.7842161

Wei ZQ, Yuan JH, Ng DWK, et al., 2016b. A survey of downlink non-orthogonal multiple access for 5G wireless communication networks. *ZTE Commun*, 14(4):17-25. https://doi.org/10.3969/j.issn.1673-5188.2016.04.003

Wei ZQ, Ng DWK, Yuan JH, et al., 2017. Optimal resource allocation for power-efficient MC-NOMA with imperfect channel state information. *IEEE Trans Commun*, 65(9):3944-3961. https://doi.org/10.1109/TCOMM.2017.2709301

Wu QQ, Li GY, Chen W, et al., 2017. An overview of sustainable green 5G networks. *IEEE Wirel Commun*, 24(4):72-80. https://doi.org/10.1109/MWC.2017.1600343

Wu W, Yin XJ, Deng P, et al., 2019. Transceiver design for downlink SWIPT NOMA systems with cooperative full-duplex relaying. *IEEE Access*, 7:33464-33472. https://doi.org/10.1109/ACCESS.2019.2904734

Wu YN, Chen XM, 2016. Robust beamforming and power splitting for secrecy wireless information and power transfer in cognitive relay networks. *IEEE Commun Lett*, 20(6):1152-1155. https://doi.org/10.1109/LCOMM.2016.2553019

Xi W, Zhou H, 2016. Enhanced CSI feedback scheme for non-orthogonal multiple access. Proc Wireless Days, p.1-3. https://doi.org/10.1109/WD.2016.7461473

Xia B, Wang JL, Xiao KX, et al., 2018. Outage performance analysis for the advanced SIC receiver in wireless NOMA systems. *IEEE Trans Veh Technol*, 67(7):6711-6715. https://doi.org/10.1109/TVT.2018.2813524

Xiao L, Li YD, Dai CH, et al., 2018. Reinforcement learning-based NOMA power allocation in the presence of smart jamming. *IEEE Trans Veh Technol*, 67(4):3377-3389. https://doi.org/10.1109/TVT.2017.2782726

Xiao Y, Hao L, Ma Z, et al., 2018. Forwarding strategy selection in dual-hop NOMA relaying systems. *IEEE Commun Lett*, 22(8):1644-1647. https://doi.org/10.1109/LCOMM.2018.2803809

Xiao ZY, Zhu LP, Choi J, et al., 2018. Joint power allocation and beamforming for non-orthogonal multiple access (NOMA) in 5G millimeter wave communications. *IEEE Trans Wirel Commun*, 17(5):2961-2974. https://doi.org/10.1109/TWC.2018.2804953

Xiao ZY, Zhu LP, Gao Z, et al., 2019. User fairness non-orthogonal multiple access (NOMA) for millimeter-wave communications with analog beamforming. *IEEE Trans Wirel Commun*, 18(7):3411-3423. https://doi.org/10.1109/TWC.2019.2913844

Xu L, Zhou Y, Wang P, et al., 2018. Max-min resource allocation for video transmission in NOMA-based cognitive wireless networks. *IEEE Trans Commun*, 66(11):5804-5813. https://doi.org/10.1109/TCOMM.2018.2854584

Xu LD, He W, Li SC, 2014. Internet of Things in industries: a survey. *IEEE Trans Ind Inform*, 10(4):2233-2243. https://doi.org/10.1109/TII.2014.2300753

Xu P, Cumanan K, 2017. Optimal power allocation scheme for non-orthogonal multiple access with α-fairness. *IEEE J Sel Areas Commun*, 35(10):2357-2369. https://doi.org/10.1109/JSAC.2017.2729780

Xu YQ, Shen C, Ding ZH, et al., 2017. Joint beamforming and power-splitting control in downlink cooperative SWIPT NOMA systems. *IEEE Trans Signal Process*, 65(18):4874-4886. https://doi.org/10.1109/TSP.2017.2715008

Xue C, Zhang Q, Li Q, et al., 2017. Joint power allocation and relay beamforming in nonorthogonal multiple access amplify-and-forward relay networks. *IEEE Trans Veh Technol*, 66(8):7558-7562. https://doi.org/10.1109/TVT.2017.2657741

Yalcin AZ, Yuksel M, Bahceci I, 2019. Downlink MU-MIMO with QoS aware transmission: precoder design and performance analysis. *IEEE Trans Wirel Commun*, 18(2):969-982. https://doi.org/10.1109/TWC.2018.2886903

Yang N, Wang LE, Geraci G, et al., 2015. Safeguarding 5G wireless communication networks using physical layer security. *IEEE Commun Mag*, 53(4):20-27. https://doi.org/10.1109/MCOM.2015.7081071

Yang Q, Wang HM, Ng DWK, et al., 2017. NOMA in downlink SDMA with limited feedback: performance analysis and optimization. *IEEE J Sel Areas Commun*, 35(10):2281-2294. https://doi.org/10.1109/JSAC.2017.2725107

Yang Z, Ding ZG, Fan PZ, et al., 2016. A general power allocation scheme to guarantee quality of service in downlink and uplink NOMA systems. *IEEE Trans Wirel Commun*, 15(11):7244-7257. https://doi.org/10.1109/TWC.2016.2599521

Yang ZH, Xu W, Pan CH, et al., 2017. On the optimality of power allocation for NOMA downlink with individual QoS constraints. *IEEE Commun Lett*, 21(7):1649-1652. https://doi.org/10.1109/LCOMM.2017.2689763

Yang ZH, Pan CH, Xu W, et al., 2018. Compressive sensing-based user clustering for downlink NOMA systems with decoding power. *IEEE Signal Process Lett*, 25(5):660-664. https://doi.org/10.1109/LSP.2018.2817181

Yu YH, Chen H, Li YH, et al., 2017a. Antenna selection in MIMO cognitive radio-inspired NOMA systems. *IEEE Commun Lett*, 21(12):2658-2661. https://doi.org/10.1109/LCOMM.2017.2750153

Yu YH, Chen H, Li YH, et al., 2017b. Antenna selection for MIMO-NOMA networks. Proc Int Conf on Communications, p.1-6. https://doi.org/10.1109/ICC.2017.7996799

Yue XW, Liu YW, Kang SL, et al., 2018a. Exploiting full/half-duplex user relaying in NOMA systems. *IEEE Trans Commun*, 66(2):560-575. https://doi.org/10.1109/TCOMM.2017.2749400

Yue XW, Liu YW, Kang SL, et al., 2018b. Spatially random relay selection for full/half-duplex cooperative NOMA networks. *IEEE Trans Commun*, 66(8):3294-3308. https://doi.org/10.1109/TCOMM.2018.2809740

Yuksel M, Erkip E, 2007. Multiple-antenna cooperative wireless systems: a diversity-multiplexing tradeoff perspective. *IEEE Trans Inform Theory*, 53(10):3371-3393. https://doi.org/10.1109/TIT.2007.904972

Zanella A, Bui N, Castellani A, 2014. Internet of Things for smart cities. *IEEE Internet Things J*, 1(1):22-32. https://doi.org/10.1109/JIOT.2014.2306328

Zaw CW, Tun YK, Hong CS, 2017. User clustering based on correlation in 5G using semidefinite programming. Proc 19th Asia-Pacific Network Operations and Management Symp, p.342-345. https://doi.org/10.1109/APNOMS.2017.8094167

Zeng M, Yadav A, Dobre OA, et al., 2017. Capacity comparison between MIMO-NOMA and MIMO-OMA with multiple users in a cluster. *IEEE J Sel Areas Commun*, 35(10):2413-2424. https://doi.org/10.1109/JSAC.2017.2725879

Zeng M, Yadav A, Dobre OA, et al., 2018. Energy-efficient power allocation for MIMO-NOMA with multiple users in a cluster. *IEEE Access*, 6:5170-5181. https://doi.org/10.1109/ACCESS.2017.2779855

Zeng M, Hao W, Dobre OA, et al., 2019. Energy-efficient power allocation in uplink mmwave massive MIMO with NOMA. *IEEE Trans Veh Technol*, 68(3):3000-3004. https://doi.org/10.1109/TVT.2019.2891062

Zhang D, Zhu ZY, Xu C, et al., 2017. Capacity analysis of NOMA with mmwave massive MIMO systems. *IEEE J Sel Areas Commun*, 35(7):1606-1618. https://doi.org/10.1109/JSAC.2017.2699059

Zhang HJ, Fang F, Cheng JL, et al., 2018. Energy-efficient resource allocation in NOMA heterogeneous networks. *IEEE Wirel Commun*, 25(2):48-53. https://doi.org/10.1109/MWC.2018.1700074

Zhang J, Andrews JG, 2010. Adaptive spatial intercell interference cancellation in multicell wireless networks. *IEEE J Sel Areas Commun*, 28(9):1455-1468. https://doi.org/10.1109/JSAC.2010.101207

Zhang L, Liu JQ, Xiao M, et al., 2017. Performance analysis and optimization in downlink NOMA systems with cooperative full-duplex relaying. *IEEE J Sel Areas Commun*, 35(10):2398-2412. https://doi.org/10.1109/JSAC.2017.2724678

Zhang NB, Wang J, Kan GX, et al., 2016. Uplink nonorthogonal multiple access in 5G systems. *IEEE Commun Lett*, 20(3):458-461. https://doi.org/10.1109/LCOMM.2016.2521374

Zhang Q, Li QZ, Qin JY, 2016. Robust beamforming for nonorthogonal multiple-access systems in MISO channels. *IEEE Trans Veh Technol*, 65(12):10231-10236. https://doi.org/10.1109/TVT.2016.2547998

Zhang SQ, Wu QQ, Xu SG, et al., 2017. Fundamental green tradeoffs: progresses, challenges, and impacts on 5G networks. *IEEE Commun Surv Tutor*, 19(1):33-56. https://doi.org/10.1109/COMST.2016.2594120

Zhang XK, Gao Q, Gong C, et al., 2017. User grouping and power allocation for NOMA visible light communication multi-cell networks. *IEEE Commun Lett*, 21(4):777-780. https://doi.org/10.1109/LCOMM.2016.2642921

Zhang Y, Yang Q, Zheng TX, et al., 2016a. Energy efficiency optimization in cognitive radio inspired non-orthogonal multiple access. Proc 27th Annual Int Symp on Personal, Indoor, and Mobile Radio Communications, p.1-6. https://doi.org/10.1109/PIMRC.2016.7794658

Zhang Y, Wang HM, Yang Q, et al., 2016b. Secrecy sum rate maximization in non-orthogonal multiple access. *IEEE Commun Lett*, 20(5):930-933. https://doi.org/10.1109/LCOMM.2016.2539162

Zheng HY, Hou SJ, Li H, et al., 2018. Power allocation and user clustering for uplink MC-NOMA in D2D underlaid cellular networks. *IEEE Wirel Commun Lett*, 7(6):1030-1033. https://doi.org/10.1109/LWC.2018.2845398

Zhong CJ, Zhang ZY, 2016. Non-orthogonal multiple access with cooperative full-duplex relaying. *IEEE Commun Lett*, 20(12):2478-2481. https://doi.org/10.1109/LCOMM.2016.2611500

Zhou FH, Chu Z, Sun HJ, et al., 2018a. Artificial noise aided secure cognitive beamforming for cooperative MISO-NOMA using SWIPT. *IEEE J Sel Areas Commun*, 36(4):918-931. https://doi.org/10.1109/JSAC.2018.2824622

Zhou FH, Wu YP, Liang YC, et al., 2018b. State of the art, taxonomy, and open issues on cognitive radio networks with NOMA. *IEEE Wirel Commun*, 25(2):100-108. https://doi.org/10.1109/MWC.2018.1700113

Zhou Y, Wong VWS, Schober R, 2018. Dynamic decode-and-forward based cooperative NOMA with spatially random users. *IEEE Trans Wirel Commun*, 17(5):3340-3356. https://doi.org/10.1109/TWC.2018.2810083

Zhu JY, Wang JH, Huang YM, et al., 2017. On optimal power allocation for downlink non-orthogonal multiple access systems. *IEEE J Sel Areas Commun*, 35(12):2744-2757. https://doi.org/10.1109/JSAC.2017.2725618

Zhu LF, Zhao HB, Liang D, et al., 2015. Mutual coupling research of multi-antenna in dual-channel balise. Proc 18th Int Conf on Intelligent Transportation Systems, p.2200-2204. https://doi.org/10.1109/ITSC.2015.355

Zhu LP, Zhang J, Xiao ZY, et al., 2018. Optimal user pairing for downlink non-orthogonal multiple access (NOMA). *IEEE Wirel Commun Lett*, 8(2):328-331. https://doi.org/10.1109/LWC.2018.2853741

Dr. Xiao-ming CHEN, corresponding author of this invited review article, received his PhD degree in Communication and Information Systems from Zhejiang University, where he is currently a professor. From January 2015 to June 2016, he was a Humboldt Research Fellow at the University of Erlangen-Nuremberg, Germany. His research interests include 5G/6G key techniques, IoT theories and techniques, and smart communications. Dr. CHEN was an editor for *IEEE Commun Lett*, and is now serving as an editor of *IEEE Trans Commun*, a corresponding expert of *Front Inform Technol Electron Eng*, and a lead guest editor of the special issue "Massive Access for 5G and Beyond" in *IEEE JSAC*. He was an exemplary reviewer for *IEEE Commun Lett* in 2014, and for *IEEE Trans Commun* from 2015 to 2018. He was a recipient of Best Paper Awards of IEEE/CIC ICCC 2018 and IEEE ICC 2019.