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:

Embracing non-orthogonal multiple access in future wireless networks*

Zhi-guo DING‡[1,2], Mai XU[3], Yan CHEN[4], Mu-gen PENG[5], H. Vincent POOR[1]

[1]*Department of Electrical Engineering, Princeton University, Princeton NJ 08544, USA*

[2]*School of Electrical and Electronic Engineering, the University of Manchester, Manchester M13 9PL, UK*

[3]*School of Electronic and Information Engineering, Beihang University, Beijing 100191, China*

[4]*Huawei Technologies Co., Ltd., Shanghai 200000, China*

[5]*Institute of Telecommunications, Beijing University of Posts and Telecommunications, Beijing 100876, China*

E-mail: z.ding@lancaster.ac.uk; MaiXu@buaa.edu.cn; bigbird.chenyan@huawei.com; pmg@bupt.edu.cn; poor@princeton.edu

Received Jan. 19, 2018; Revision accepted Mar. 9, 2018; Crosschecked Mar. 25, 2018

Abstract: This paper provides a comprehensive survey of the impact of the emerging communication technique, non-orthogonal multiple access (NOMA), on future wireless networks. Particularly, how the NOMA principle affects the design of the generation multiple access techniques is introduced first. Then the applications of NOMA to other advanced communication techniques, such as wireless caching, multiple-input multiple-output techniques, millimeter-wave communications, and cooperative relaying, are discussed. The impact of NOMA on communication systems beyond cellular networks is also illustrated, through the examples of digital TV, satellite communications, vehicular networks, and visible light communications. Finally, the study is concluded with a discussion of important research challenges and promising future directions in NOMA.

Key words: Non-orthogonal multiple access (NOMA); Wireless caching; Multiple-input multiple-output (MIMO) NOMA; Cooperative NOMA; Millimeter-wave networks; Visible light communications (VLC)

https://doi.org/10.1631/FITEE.1800051 **CLC number:** TN92

1 Introduction

Unlike wireline communications, the broadcast nature of wireless communications renders wireless transmission particularly prone to interference (Proakis, 2000). As a result, the use of orthogonal signalling, which provides a simple way to avoid co-channel interference, has been a dominant approach for most multiple access techniques used by the previous generations of mobile networks. For example, frequency division multiple access (FDMA) was used in the first generation of mobile networks by dividing the frequency domain into non-overlapping frequency channels. These orthogonal frequency channels were then exclusively allocated to users, which avoids multiple access interference; i.e., each user solely occupies a frequency channel and hence does not cause co-channel interference to others. Similar to the first generation, the following generations of mobile systems have also employed multiple access techniques based on the same idea that orthogonal resource blocks in frequency/time/code domains are allocated to users separately.

However, from the perspective of information theory, it is well known that the use of such orthogonal multiple access (OMA) approaches is not optimal in terms of spectral efficiency (Verdu, 1998; Cover and Thomas, 2006). Taking multi-user up-

‡ Corresponding author

* Project supported by the UK EPSRC (No. EP/N005597/1), the H2020-MSCA-RISE-2015 (No. 690750), the National Natural Science Foundation of China (No. 61728101), and the U.S. National Science Foundation (Nos. CNS-1702808 and ECCS-1647198)

ORCID: Zhi-guo DING, http://orcid.org/0000-0001-5280-384X

link transmission as an example, which is termed the 'multiple access channel' in information theory, the rate region achieved by an orthogonal multiple access approach is only a part of the capacity region of the multiple access channel, and this capacity region can be achieved if users are allowed to transmit at the same time/frequency/code (Cover and Thomas, 2006). While this performance loss of OMA has been known for more than 50 years, the OMA approaches are commonly used, because the implementation of those multiple access techniques based on non-orthogonality relies on the use of sophisticated transceiver designs. These designs typically result in high computational complexity and implementation costs, and hence could not be supported in the previous generations of mobile systems.

Starting from 2013, the telecommunication industry has been considering removing the orthogonality in the design of multiple access techniques for the next generation of mobile networks (NTT Docomo Inc., 2014; Huawei Inc., 2015; tech$^{\mathrm{UK}}$, 2015). Meanwhile, various research efforts have also been devoted to the design of new types of multiple access techniques based on the idea of spectrum sharing and serving multiple users in the same orthogonal resource block, which have been generally termed 'non-orthogonal multiple access (NOMA)' (Wei et al., 2016; Ding et al., 2017a, 2017f). This interest in NOMA is mainly due to the following three reasons:

1. Thanks to Moore's law, the computing power of devices in mobile networks has been significantly improved in recent years; e.g., smart phones in use today are as powerful as computers and are capable of high performance computing. This increase of processing power is crucial for the implementation of NOMA. For example, many forms of NOMA require the receivers to carry out successive interference cancellation (SIC), a step which has been conventionally believed infeasible at the user side. Recently a NOMA chipset-embedded device has been developed to implement SIC in smartphones (NTT Docomo Inc., 2017).

2. NOMA has been proposed at a time when the fifth generation (5G) networks are envisioned to not only support conventional voice and data services, but also provide Internet of Things (IoT) functionalities. Recall that a key feature of IoT is that the number of devices to be connected can be massive, and hence realizing massive connectivity is important to support IoT in 5G. However, conventional OMA schemes cannot realize massive connectivity straightforwardly. Take time division multiple access (TDMA) as an example. If TDMA is used to support massive connectivity, a short time duration, e.g., one millisecond, needs to be further divided into a huge number of time slots, and hence the duration of each time slot will be very short, which increases the implementation costs. Note that the use of FDMA for massive connectivity also results in a situation that adjacent frequency channels are too close, which can cause severe inter-channel interference. The use of the NOMA principle provides a more flexible way of supporting massive connectivity.

3. For future wireless networks, devices and users to be connected have diverse quality of service (QoS) requirements, to which the use of OMA is not appropriate (Ding et al., 2016d). For example, consider a scenario in which there are 10 sensors that need to be served with low data rates only, and one broadband user. The use of OMA, such as orthogonal frequency division multiple access (OFDMA), means that each sensor is allowed to solely occupy one resource block, such as one OFDM subcarrier with 10 MHz. This is a waste of valuable spectrum since sensors are given more bandwidth than they need, but the broadband user might not have enough. A more spectrally efficient way is to encourage spectrum sharing, by implementing NOMA and integrating these sensors and the broadband user into a single subcarrier.

We provide a survey of the impact of the NOMA principle on wireless communications, from the following four perspectives (Fig. 1):

1. How the NOMA principle is used to affect the design of multiple access techniques for future networks is considered. In particular, the general principle of NOMA is first discussed. Then practical forms of NOMA using a single resource block are introduced and various designs of NOMA schemes using multiple resource blocks are also described. It is important to point out that no multiple access technique, including NOMA, is perfect. This is the reason why the bandwidth resource blocks obtained from other types of OMA are used for the implementation of NOMA.

2. The impact of the NOMA principle on various advanced communication technologies, such as millimeter-wave (mmWave) transmission,

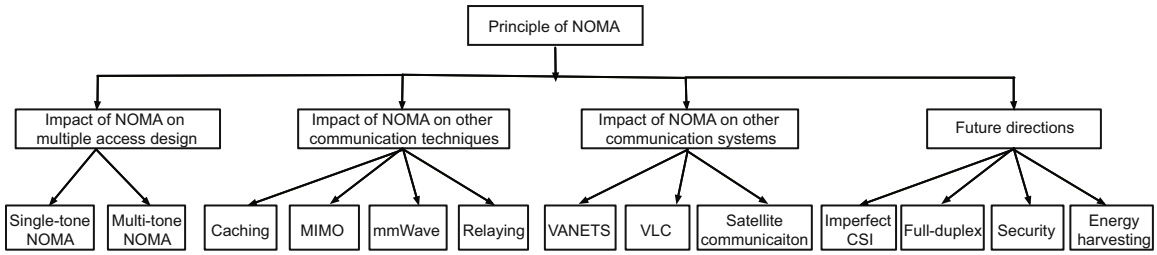

Fig. 1 An illustration for the structure of the study

multiple-input multiple-output (MIMO) techniques, and cooperative communications, is discussed. The spectral efficiency of these advanced communication technologies can be significantly improved with the application of NOMA. Furthermore, many features of these advanced communication techniques can be efficiently used to facilitate the implementation of NOMA for improving system performance.

3. The NOMA principle is shown to be useful to many communication scenarios beyond cellular networks, although the concept of NOMA was originally designed for cellular networks. For example, in addition to radio frequency communication networks, the NOMA principle has been shown to be particularly useful to the design of visible light communication (VLC) networks. Another example is that the NOMA principle can be straightforwardly applied to scenarios beyond mobile communications, such as terrestrial TV broadcasting and satellite communications. These discussions will illustrate that NOMA not only brings new opportunities for the design of future multiple access techniques, but also has the capability to shape future communication networks.

4. Important directions for future research on NOMA are outlined and discussed. In particular, the challenges for the implementation of NOMA with imperfect channel state information (CSI) are described. The potential for the applications of NOMA to physical layer security, full duplex communication systems, and radio frequency and VLC based energy harvesting is also illustrated and discussed.

2 A paradigm shift in designing multiple access techniques

2.1 General principles of NOMA

The essential principle of NOMA is to encourage spectrum sharing among multiple users, instead of allowing them to solely occupy orthogonal resource blocks. The basic idea of NOMA can be clearly illustrated using the example of two-user downlink power-domain NOMA (Saito et al., 2013; Ding et al., 2014; Choi, 2016a). As its name suggests, power-domain NOMA uses the power domain for multiple access. If power-domain NOMA is used, the base station first superimposes the users' signals and broadcasts this mixture to all the users. As a result, all the users are served at the same time/frequency/code, but with different power levels. These power levels are decided by the superposition coefficients, also termed 'power allocation coefficients'. Power allocation of power-domain NOMA is different from conventional power allocation. Without loss of generality, we consider a two-user downlink scenario, in which the users are to receive different messages from their base station. Particularly, more power is allocated to the user with poorer channel conditions. The reason for this type of power allocation is to ensure user fairness, since NOMA is a multiple access technique and needs to ensure that all the users are served. Assigning more power to the user with stronger channel conditions might improve the throughput, but can cause the user with poorer channel conditions to become disconnected.

The receivers of power-domain NOMA have different detection strategies, according to their channel conditions. Particularly, the user with poorer channel conditions treats its partner's information as noise, and directly decodes its own information, which is feasible since its own message enjoys a higher power level than its partner's message. On the other hand, the user with stronger channel conditions will have to decode its parter's information first, before decoding its own, a procedure known as SIC (Nonaka et al., 2014). The reason to use SIC at the user with stronger channel conditions is due to the use of power-domain NOMA power allocation; i.e., its

own message was buried underneath its partner's information. The benefit of NOMA can be easily illustrated by considering an extreme case, in which the user with poorer channel conditions experiences deep fading. In this case, the use of conventional OMA, such as OFDMA, is inefficient, since the subcarrier allocated to the weak user is wasted. By using NOMA, the bandwidth solely occupied by the weak user in the OMA mode can be released and used by other users, which significantly improves the spectral efficiency.

2.2 Implementing NOMA at a single bandwidth resource block

When multiple users are to share a single bandwidth resource block, NOMA can be implemented by simply using power-domain NOMA, as explained in the previous subsection. The key idea of power-domain NOMA is to allocate more power to users with weaker channel conditions. However, how much power should be allocated to these users is not rigorously defined, which leads to an issue that power-domain NOMA cannot strictly guarantee the users' QoS requirements. In addition, power-domain NOMA cannot be applied to the scenario in which users have similar channel conditions. These become the motivations for another form of NOMA, termed 'cognitive radio (CR) inspired NOMA' (Ding et al., 2016c; Yang et al., 2016a; Mitra and Bhatia, 2017). CR-NOMA treats NOMA as a special case of cognitive radio networks. Again take the two-user downlink case as an example. The weak user can be viewed as a primary user in cognitive radio networks, and the use of OMA is equivalent to a situation without any spectrum sharing; i.e., the weak user solely occupies the bandwidth. NOMA introduces spectrum sharing, in which a strong user, viewed as a secondary user in cognitive radio networks, is introduced to the system. Since the secondary user has a strong connection to the base station, it can significantly improve the overall system throughput. By using this synergy between NOMA and cognitive radio networks, a new form of NOMA, CR-inspired NOMA, can be developed (Ding et al., 2016c; Yang et al., 2016a).

The key difference between power-domain NOMA and CR-NOMA lies in two aspects: how the users are ordered and how the transmission power is allocated among the users. Particularly, CR-NOMA orders users according to their QoS requirements, instead of their channel conditions. The CR-NOMA power allocation policy is to first provide sufficient power to the users with strict QoS requirements, and the remaining power, if there is any left, is allocated to the users that can be served opportunistically. The performance gain of CR-NOMA over OMA and power-domain NOMA can be illustrated by a simple two-user downlink example, with the following assumptions:

1. User 1 needs to be served in a timely manner, but its targeted data rate is low. Without loss of generality, assume the target rate is 1 bit/(Hz·s). In practice, this type of users can be wireless healthcare devices or wireless sensors for disaster management.

2. User 2 is delay-tolerant and can be served opportunistically, e.g., a user having a data downloading task in the background for system updates.

3. Both the users have the same channel gains, which are assumed to be 1 for the purpose of illustration.

Since the channel gains are the same, for this scenario, the sum rate offered by power-domain NOMA is the same as OMA; i.e., power-domain NOMA is not applicable to this scenario. When the transmitted signal-to-noise ratio (SNR) ρ is high, i.e., $\rho \to \infty$, the sum rates achieved by OMA and CR-NOMA can be approximated as follows:

1. Each user's rate in OMA can be approximated as $\frac{1}{2}\log\rho$ bits/(Hz·s). However, user 1 only needs to be served with a rate of 1 bit/(Hz·s). So, the sum rate of OMA can be approximated as $(1+\frac{1}{2}\log\rho)$ bits/(Hz·s).

2. At a high SNR, a very small amount of power needs to be consumed to guarantee the low targeted data rate of user 1. Therefore, the sum rate of CR-NOMA can be approximated as $(1+\log\rho)$ bits/(Hz·s), which is much larger than that of OMA.

CR-NOMA also has other features that are different from power-domain NOMA. For example, the outage probability of a user in power-domain NOMA is determined only by its own channel condition, not by the other users' channels. However, in CR-NOMA, the outage performance of the users that are served opportunistically is not only related to their own channel conditions, but also determined by the channel quality of the other users. This is because CR-NOMA first serves users with strict QoS require-

ments, meaning that the power available to opportunistic users is determined by the channel conditions of the users with strict QoS requirements.

2.3 Implementing NOMA with multiple bandwidth resource blocks

2.3.1 Hybrid NOMA

Hybrid NOMA refers to a type of NOMA implementation in which each user is allowed to use multiple bandwidth resource blocks simultaneously and each resource block is to accommodate multiple users (Ding et al., 2016c). The key motivation for hybrid NOMA is to reduce the complexity for the implementation of NOMA. For example, consider a scenario, in which there are 100 users in a cell. If all the users are grouped into a single group for the implementation of NOMA, the best user has to decode the remaining 99 users' signals before decoding its own, which is obviously not feasible. Hybrid NOMA provides a low-complexity alternative for the implementation of NOMA. To be consistent with the exiting literature about hybrid NOMA, we use OFDMA subcarriers as examples of bandwidth resource blocks, given the fact that OFDMA will be used in 5G. Again take the 100-user case as an example. Hybrid NOMA can divide these users into 20 groups with five users in each group. Different OFDMA subcarriers are allocated to different groups to avoid inter-group interference. Within each group, NOMA can be applied to serve five users on the same subcarrier, which significantly reduces the system complexity.

Note that hybrid multiple access techniques have already been used in the previous generations of mobile networks. For example, in global system for mobile communication (GSM) systems, eight time slots created by TDMA are not sufficient to support a system with a large number of users, which motivates the combination of TDMA and FDMA in GSM. In the third generation (3G) mobile system, frequency division duplex is combined with code division multiple access (CDMA) to provide sufficient connections with reasonable reception reliability to multiple users. The fourth generation (4G) mobile network also employs hybrid multiple access, where TDMA and OFDMA are efficiently combined together. Following the same rationale, it is expected that NOMA is also to be implemented in this hybrid manner in future wireless networks.

2.3.2 User grouping

A key step to design hybrid NOMA is user grouping, since the overall system performance depends on which user is grouped with whom at which subcarrier (Yakou and Higuchi, 2015; Sun et al., 2017; Zhang X et al., 2017). Initial studies about user grouping have drawn some interesting conclusions, as discussed below (Ding et al., 2016c). Provided that power-domain NOMA is implemented and users are grouped according to their channel conditions, one important conclusion is that users with different channel conditions can have completely different experiences. Particularly, a user with strong channel conditions benefits the implementation of NOMA, since this user's data rate in NOMA is very likely to be higher than that in OMA. On the other hand, a user with poor channel conditions may suffer some data rate loss, compared with the case with OMA, as it experiences strong co-channel interference caused by its partner. Another important conclusion is that, if CR-NOMA is used, the QoS requirements of the primary users can be strictly guaranteed, but the performance achieved by those secondary users depends largely on the channel conditions of the primary users, as discussed in the previous subsection on CR-NOMA.

Based on these insights, various user grouping algorithms have been developed in hybrid NOMA networks. It is worth pointing out that finding optimal user pairings for hybrid NOMA is not a trivial problem to solve, as it is essentially an integer programming problem. Furthermore, the user pairing issue is coupled with other optimization problems, such as power allocation and subcarrier allocation, which makes the overall system optimization very challenging. In Sun et al. (2017), the monotonic optimization tool has been applied to hybrid NOMA for joint user grouping and power allocation. The benefit of using this tool is to ensure that an optimal solution for the non-convex mixed integer optimization problem can be found. While the computational complexity of the monotonic optimization tool is high, the use of this tool is still important, as it provides a useful benchmark for low-complexity suboptimal solutions. Also, note that other optimization tools other than monotonic optimization, such as branch-and-bound algorithms and machine learning methods, can also be applied to the addressed

the problem (Boyd and Vandenberghe, 2004).

2.3.3 Practical forms of hybrid NOMA

Because of its low complexity and superior spectral efficiency, the industry has developed various forms of hybrid NOMA. One of the most well-known hybrid NOMA is sparse code multiple access (SCMA) (Nikopour and Baligh, 2013; Taherzadeh et al., 2014). The key advantage of SCMA is overloading, where the number of subcarriers is smaller than the number of the supported users, which is important for realizing massive connectivity. The sparsity feature of SCMA is due to the requirement that each user is allowed to use a very small number of subcarriers. This sparsity feature is important to reduce the system complexity since the number of users occupying the same subcarrier becomes small. Compared with power-domain NOMA, SCMA exhibits two differences, one at the transmitter and the other at the receiver, as explained as follows:

1. Unlike power-domain NOMA, SCMA requires the use of multi-dimensional coding at the transmitter, and the reason to use this coding is explained below. In SCMA, each user can use multiple subcarriers to transmit a single data stream, a feature also termed 'low-density spreading', and how subcarriers are allocated to a user is determined by the factor graph matrix (Cai et al., 2016; Yang et al., 2016b). One option to use the multiple subcarriers is to generate multiple identical copies of the user's data stream and to send them over the multiple subcarriers, as done in low-density spreading. However, SCMA adopts a more efficient way that generates correlated copies of the data stream and sends these copies over the subcarriers.

2. At the receiver, SCMA uses the message passing algorithm (MPA) instead of SIC, which is due to the following reason. If a user's information spread over multiple subcarriers is independently coded, SIC can be applied to decode the user's information at each subcarrier individually and then maximum ratio combining can be used to combine the decoded information from different subcarriers. However, owing to the use of multi-dimensional coding, one user's messages transmitted over different subcarriers are correlated, from which the MPA yields better performance than SIC (Yu et al., 2016, 2017).

There are other types of hybrid NOMA, for example, pattern division multiple access (PDMA) (Chen S et al., 2017a). A user's information is spread over multiple subcarriers, similar to SCMA, but the sparsity constraint of SCMA is removed; i.e., one user might use many subcarriers in PDMA. Since there are fewer constraints for subcarrier allocation in PDMA, there are more degrees of freedom for system design, which can be used to improve the system performance albeit at a price of increased complexity.

3 Applying NOMA to other advanced communication technologies

The NOMA principle not only brings changes for the design of the next generation multiple access techniques, but also has an important impact on the design of other advanced communication technologies.

3.1 NOMA-assisted wireless caching

The key idea of wireless caching is to proactively push popular content files to local caching infrastructure, e.g., local content servers or other users in the device-to-device (D2D) caching case (Golrezaei et al., 2013; Bastug et al., 2014). As a result, when the users request these files, they do not need to communicate directly with the base station, but simply fetch the files from their local content servers or D2D helpers. The benefit of wireless caching can be illustrated by the following example. Consider that there are 100 users requesting different files. Without wireless caching, 100 resource blocks need to be consumed to accommodate these users' requests. However, provided that these files have been previously cached by the local content servers, only one resource block is needed to serve these 100 users. The reason for this is that the content servers can help their associated users locally and that the use of short-range communications ensures that all the transmission by the content servers can be carried out simultaneously.

Conventional wireless caching assumes that content pushing is carried out by using off-peak hours, during which a lot of the spectrum is idle and can be used for content pushing (Golrezaei et al., 2013; Maddah-Ali and Niesen, 2014; Chen and Kountouris, 2016; Xu and Tao, 2017). This assumption is valid if the popularity of the content files varies slowly. Typical examples for this type of content are software updates, popular movies, and TV streaming.

However, many other types of content, such as up-to-date sporting event news and sale pricing information, exhibit a fast time-varying feature and need to be updated frequently. For these types of content, the assumption that using off-peak hours for caching is not applicable, since the files cached during off-peak hours might become outdated during peak hours. The application of the NOMA principle can bring some fundamental changes for the design of wireless caching, as illustrated below (Ding et al., 2017b).

When off-peak hours cannot be used, content pushing has to be carried out during peak hours. To keep the files cached at the local content servers frequently updated, a short-time duration needs to be periodically used for content pushing. This periodically used duration has to be short since not all the pushed files are useful for users and spending a large amount of time for content pushing will reduce the spectral efficiency of wireless caching. If OMA-based content pushing is used, this duration will be further divided into small time slots, and the base station will push one file to a single content server during each time slot. If the number of the content servers is large, some content servers might not obtain any file pushed from the base station, which is the drawback of OMA-based content pushing. If the NOMA principle is applied, the base station can superimpose multiple content files that are intended to different content servers, and uses one time slot to serve multiple content servers. As a result, NOMA-based caching is more suitable to meet the constraint that limited bandwidth resources are reserved for content pushing. Similarly, the concept of NOMA can be applied to the content delivery stage. Recall that the purpose of the content delivery stage is to ask the content servers to serve their associated users, if these users' requested files can be found locally. The drawback of OMA-based content delivery is that at each time, a content server can serve one user only. However, for many high-density wireless networks used in airports or stadiums, it is very likely that one content server will serve more than one user. The use of NOMA can ensure that multiple users can be connected to the same content server, which improves the latency of wireless caching, since users do not have to wait for a long time to be served.

Another NOMA-assisted wireless caching scheme is to opportunistically carry out content pushing during the content delivery stage. In conventional wireless caching, the stages for content pushing and content delivery are strictly separated; i.e., time slots during the content delivery stage cannot be used for content pushing. However, if the time duration between two adjacent content pushing stages is large, the content files at the local content servers cannot be frequently updated. By using the NOMA principle, this drawback of conventional wireless caching can be avoided. Particularly, some time slots during the content delivery stage can be identified as opportunities for content pushing. For example, during some time slots in the content delivery stage, users make requests to be served, but their requested files cannot be found in the caches of the local content servers. Conventionally this type of event is viewed as non-ideal since the base station has to serve these users directly and hence the spectral efficiency of wireless caching is reduced. With the application of NOMA, the base station can superimpose two types of signals, one to be delivered to the users directly and the other to be pushed to the content servers. As a result, the base station does not have to wait until the next content pushing stage to push files to the content servers, and the files stored in the local caches can be frequently updated.

3.2 MIMO-NOMA

The NOMA principle has a significant impact on the design of MIMO technologies. Particularly, spatial directions can also be viewed as a type of bandwidth resource blocks. Conventional MIMO techniques, such as zero forcing, prefer to serve a single user at one of multiple orthogonal spatial directions, whereas the use of NOMA ensures that more users can be connected at a single spatial direction (Foschini and Gans, 1998). In the following, general principles of MIMO-NOMA are discussed first, and then some practical designs are introduced.

3.2.1 General principles

Unlike single-input single-output (SISO) NOMA, it is very challenging to identify the optimality of MIMO-NOMA. Without loss of generality, we focus mainly on downlink NOMA below. In Xu et al. (2015), the relationship between the rate region achieved by SISO-NOMA and the

capacity region of broadcast channels has been clearly illustrated. However, little is known about optimal MIMO-NOMA, partially because the capacity region for general broadcast channels is still unknown. Note that dirty paper coding (DPC) has been well accepted as a reasonable benchmark given its capability to approach an upper bound on the capacity region. Therefore, it is of interest to study the comparison between NOMA and DPC.

In Chen et al. (2016a, 2016b), a condition for NOMA to realize the same performance as DPC, termed the quasi-degradation criterion, is established, for the multi-input single-output (MISO) scenario in which the base station has multiple antennas and each user has a single antenna. Provided that users' channels satisfy the quasi-degradation criterion, the use of NOMA yields the same performance as DPC, but the complexity of NOMA is linearly proportional to the number of users, much smaller than that of DPC. The following two examples are provided to illustrate the key idea of the quasi-degradation criterion:

1. When users' vectors have the same directions but different magnitudes, the quasi-degradation criterion is satisfied. The optimality of NOMA in this scenario is intuitive, since a beamforming vector good for one user is also good for the others; i.e., users are located in the same spatial direction and hence can be served using a single beam.

2. The quasi-degradation criterion cannot be satisfied if users have orthogonal channels. It is also intuitive that NOMA cannot be applied to this scenario since one user's beam is useless to the others due to the orthogonality of the users' channels.

However, the quasi-degradation criterion hs been developed for MISO-NOMA only, and its extension to general MIMO-NOMA is still unknown.

3.2.2 Practical designs of MIMO-NOMA

Even though the optimality of MIMO-NOMA is still unknown, it is worth developing practical MIMO-NOMA designs, with the aim that they can outperform MIMO-OMA. One popular way of designing MIMO-NOMA is to ask the base station to generate many non-orthogonal beams, where a single user is accommodated by one beam (Hanif et al., 2016; Zeng et al., 2017). Since the generated beams are non-orthogonal, overloading can be supported by this type of MIMO-NOMA; i.e., the number of supported users is larger than the number of antennas at the base station. A key challenge for this type of MIMO-NOMA is how to order users according to their channel conditions, since channels are in the form of vectors or matrices. The existing studies in Choi (2016b) and Hanif et al. (2016) have shown that the use of path loss for user ordering can ensure a reasonable performance gain over OMA.

Another way of designing MIMO-NOMA is to decompose MIMO-NOMA into SISO-NOMA by carefully designing precoding and detection matrices (Ding et al., 2016a, 2016b). Particularly, the spatial degrees of freedom are first used to create some orthogonal beams by using conventional MIMO techniques, and then the NOMA principle is applied to ensure that multiple users can be served by each of the generated beams. The benefit of this type of MIMO-NOMA is that there is no need to directly order users' channel vectors/matrices, since after converting MIMO-NOMA to SISO-NOMA, the effective channel gains are in the form of scalars, instead of vectors or matrices. In addition, this type of NOMA facilitates the implementation of hybrid NOMA, and it is also applicable to both uplink and downlink transmissions. Furthermore, this type of MIMO-NOMA design is particularly suitable for massive MIMO scenarios, where the users sharing the same channel correlation matrix can be grouped together and served by the same beam (Ding and Poor, 2016; Zhang D et al., 2017).

3.3 MmWave-NOMA

With the rapid growth of traffic demand, the radio spectrum below 6 GHz used by conventional wireless networks has become too crowded, which has motivated the recent interest in using the less occupied mmWave spectrum (Heath et al., 2016; Kulkarni et al., 2016). Note that the motivation to use NOMA is exactly the same as that for using mmWave, with NOMA aiming to improve the spectral efficiency within the available bandwidth. Obviously mmWave communications and NOMA are not conflicting but are complementary to each other. On the one hand, the mmWave bands are not free of charge, but can be very expensive according to the lessons learned from 3G/4G spectrum auctions, which motivates the use of NOMA in mmWave communications as a cost-effective measure. On the other hand, even if mmWave bands turn out to be

much less expensive than the lower-frequency ones, the tremendous increase in the number of mobile devices and the types of bandwidth demanding services, such as ultra-high definition video streaming and online interactive games, will soon place a strict requirement on how efficiently the mmWave bands are used, which also motivates the use of NOMA in mmWave networks.

In addition to the aforementioned motivations, the application of NOMA can efficiently use some features of mmWave transmission, and hence significantly improve the spectral efficiency of mmWave communications. For example, one of the key features of mmWave communications is that mmWave transmission is highly directional. In conventional wireless communications using frequencies lower than 6 GHz, the channels of two receivers that are spaced more than half of the wavelength apart can be assumed to be independent, due to multipath fading; i.e., the number of paths between a transmitter and a receiver can approach infinity in a rich scattering environment. However, paths in mmWave transmission are very few, and the line-of-sight path is dominant, which means that two users' channels can be highly correlated, even if the distance between the two users is large. According to the quasi-degradation criterion (Chen et al., 2016a), correlated users' channels are ideal for the application of NOMA, in which a single beam generated by the base station can accommodate both users. The benefit for this type of mmWave-NOMA can be explained by using the following example (Ding et al., 2017d; Zhang Z et al., 2017b). Consider that there are eight single-antenna users and that the base station has four antennas only. The use of conventional zero forcing can only ensure that the four users are simultaneously served by the base station. By applying mmWave-NOMA and exploring the channel correlation, all the users might be supported at the same time. Furthermore, conventional zero forcing results in poor reception reliability if users' channels are correlated, since it tries to create two orthogonal beams for these users. However, spatial degrees of freedom can be more efficiently used in mmWave-NOMA, by accommodating users with correlated channels in a single orthogonal direction and handling the intra-beam interference.

Another example for the features of mmWave communications to facilitate the application of NOMA is the use of finite-resolution analog beamforming (FRAB) (Alkhateeb et al., 2016; Gao et al., 2017). In particular, FRAB is a special case of analog beamforming, in which the phases of the transmitted signals are changed, but their amplitudes are kept the same. The reason to use analog beamforming is mainly due to the high cost of radio frequency chains, where changing the signal amplitudes can be much more expensive than changing the signal phases. In practice, the signal phases are changed by using phase shifters, and the number of phase shifts supported by practical circuits is limited. For example, if a perfect analog beamformer requires a shift of 1×10^{-5} degree, most likely it cannot be supported in practice. It is worth pointing out that FRAB is not only applicable to mmWave communications, but also commonly used in massive MIMO systems. While the use of FRAB can significantly reduce the hardware cost, it is well known that this type of imperfect beamforming causes performance degradation, since these generated beamforming vectors are not perfectly aligned with the users' channels. However, the feature of FRAB that beams are not aligned with users' channels can be used to facilitate the implementation of NOMA, as illustrated in the following example (Ding et al., 2017c). Consider that there are two single-antenna users and that the base station has two antennas. Furthermore, assume that the users have orthogonal channels. Using conventional zero forcing techniques, the base station can serve the two users simultaneously, which consumes all the degrees of freedom at the base station. It is preferable to apply NOMA to this scenario. So the two users can be grouped and served by one beam, which saves degrees of freedom at the base station and provides the possibility to serve additional users. According to the quasi-degradation criterion, the application of NOMA to this scenario is not possible, since the users have orthogonal channels. However, if FRAB is used to generate beams, it is possible that the two users with the orthogonal channels prefer the same FRAB vector, particularly when the resolution of FRAB is low. As a result, the base station can use a single beam to serve the two users and hence additional beams can be generated to serve more users, which is not possible if perfect beamforming is used.

3.4 Cooperative NOMA

The existing cooperative NOMA schemes can be divided into two types. The first one is to seek the

opportunity for cooperation by asking one NOMA user to help the others (Ding et al., 2015; Lv et al., 2017; Wei et al., 2017). This type of cooperative NOMA is motivated by the fact that the implementation of NOMA can degrade some NOMA users' performance. As discussed in the subsection for CR-NOMA, a far user can be viewed as a primary user, and NOMA can potentially reduce its reception reliability since an additional user is introduced to the system. The key idea of the first type of cooperative NOMA is to recruit the users with strong channel conditions as relays and help those with poor channel conditions. It is worth pointing out that the SIC feature of the NOMA receivers facilitates the cooperation among NOMA users. In particular, users with strong channel conditions have to first decode the signals to the users with poor channel conditions. As a result, the information for the users with poor channel conditions becomes available to the strong users after carrying out SIC. Therefore, it is natural to use these strong NOMA users as relays, where no extra time slot is needed to deliver the weak users' information to the strong ones.

One drawback of the first type of cooperative NOMA is its limited diversity gain, since it relies on the cooperation among the active users but the number of active users in practice might be small. The second type of cooperative NOMA avoids this drawback and seeks help from dedicated relays which assist a base station to deliver the information to its users (Kim and Lee, 2015; Luo and Teh, 2017). Because inactive users in a network are much more numerous than the active ones, a higher diversity gain can be achieved by using these inactive users as dedicated relays, compared with the case that only the cooperation among the NOMA users is employed. In addition, using dedicated relays can be particularly useful if the base station does not have direct links with the NOMA users; i.e., the NOMA users are located close to the cell edge. In this case, employing cooperative NOMA can ensure that the users' information can be delivered from the relay to the users more spectrally efficiently than by using cooperative OMA, since one NOMA broadcast by a relay can help multiple users.

Note that the network topology has a great impact on the design of efficient cooperative NOMA. For example, when cooperation among NOMA users is used, some users that have strong connections to the base station but not to the weak users should not be employed as relays, as short communications for relay transmission become impossible and hence extra bandwidth resources are needed by these strong users to reach the weak users. When dedicated relays are used, different designs of cooperative NOMA can be developed depending on whether the users have direct links with the base station and which users need the help from the relay. Furthermore, considerable research effort has been devoted to a particular network topology, in which multiple relays are available for cooperative NOMA. While distributed beamforming can be used to exploit all the available relays, the coordination among these relays, such as time and phase synchronization, can consume substantial system overhead. As a result, relay selection, i.e., selecting a single relay for the cooperation, is preferred in practice. Depending on whether amplify-and-forward or decode-and-forward is used at the relays, various relay selection schemes have been developed (Ding et al., 2016e; Yang et al., 2017). The max-min relay selection strategy which has been proved to be optimal in conventional cooperative networks is no longer optimal in cooperative NOMA. The main reason for this is that the max-min criterion is to select a relay whose incoming and outgoing channels are most balanced; however, these incoming and outgoing channels are not equally important in cooperative NOMA. As shown in Ding et al. (2016e) and Yang et al. (2017), various relay selection strategies have been developed and shown to outperform the max-min selection scheme.

4 Applications of NOMA beyond cellular networks

4.1 Vehicular ad hoc networks

Vehicular ad hoc networks (VANETs) are envisioned to provide important applications related to road safety, data sharing among vehicles, intelligent transportation, etc., and also to support connected and autonomous vehicles and systems (Molina-Masegosa and Gozalvez, 2017). Originally, only vehicle-to-infrastructure (V2I) and vehicle-to-vehicle (V2V) communications were considered in VANET, but recently other types of communications, such as vehicle-to-pedestrian (V2P), vehicle-

to-device (V2D), and vehicle-to-grid (V2G), have also been considered, which leads to the general term 'vehicle-to-everything (V2X) communications' (Chen S et al., 2017b; Chen Y et al., 2017).

The key feature of V2X communications is short duration for the communication connection, which can be illustrated by using V2I communications as an example (Ho et al., 2011). V2I communications refer to the scenario where the infrastructure, such as a roadside unit or a base station, communicates with vehicles. For example, the infrastructure gathers the local information from vehicles, similar to uplink transmission in conventional cellular networks. In addition, via the downlink, the infrastructure frequently disseminates global traffic and road information to the vehicles and may also need to provide certain suggestions and instructions for real-time motion planning to autonomous vehicles. Because vehicles are moving at high speeds, the connection period between a vehicle and the infrastructure can be very short, which imposes a challenge for reliable communications over V2I channels. Connecting a large number of devices within this short duration makes the problem even more difficult.

Compared with OMA-based transmission strategies, the NOMA principle is more flexible to provide timely and massive connections, realize dynamic resource allocation, and meet users' diverse QoS requirements (Ding et al., 2016d). For example, suppose that two vehicles need to be connected with the same infrastructure, where one needs to be connected to receive real-time content and the other can be served opportunistically as it requires non-real-time servies. The use of NOMA can ensure that both the users have access to all the bandwidth resources, such as the short connection duration and the spectrum, which is particularly important to the VANET scenario. Furthermore, the transmission power allocated to the users can be flexibly designed to guarantee the QoS requirement of real-time service, while the infrastructure opportunistically delivers the non-real-time files to the other users.

Handover is another key challenge for VANETs, since a vehicle with high mobility can travel through multiple cells covered by different roadside base stations in a short period of time. Network MIMO, such as coordinated multipoint (CoMP) and cloud radio access networks (CRANs), has been recognized as an efficient method to combat the handover issue, as one vehicle is connected to multiple base stations and hence disconnection can be avoided. NOMA can improve the spectral efficiency of network MIMO. For example, while two base stations serve one user simultaneously, they cannot be accessed by other users in conventional network MIMO. However, by using NOMA, each base station can schedule additional users that are close to the base station. This strategy is particularly important to VANETs, since more users can be connected during the short connection duration (Di et al., 2017).

4.2 Visible light communications

Similar to mmWave communications, visible light communications (VLC) are motivated by the insufficient bandwidth resources below 6 GHz reserved for wireless communications, and it is preferable to use those less occupied high frequency bands (Komine and Nakagawa, 2004). Particularly, VLC uses visible light whose frequency is between 400 and 800 THz for communications. Note that acquiring more bandwidth, i.e., using mmWave and VLC, does not conflict with the goal of improving the spectral efficiency, i.e., using NOMA (Yin et al., 2016). On the contrary, how to efficiently use the spectrum is important even if there are plenty of new bandwidth resources obtained, in order to support emerging broadband services, as discussed in the subsection for mmWave-NOMA.

Similar to mmWave transmission, VLC exhibits some features that facilitate the implementation of NOMA (Zhang X et al., 2017). For example, channels for the scenario using frequencies lower than 6 GHz can suffer fast-time-varying multipath fading, which makes the design of NOMA challenging. This is because the important components of NOMA transceivers, such as SIC, MPA, and NOMA power allocation, require knowledge of CSI. Imperfect CSI can significantly degrade the performance of NOMA. However, in the context of VLC, the channels can be viewed as static. This is because VLC relies mainly on the line-of-sight path, and those effects important to conventional radio frequency systems, such as reflection and diffusion, can be ignored in VLC. With these static channels, the implementation of NOMA becomes more straightforward than for systems using radio frequencies.

In addition, it is well known that the performance gain of NOMA over OMA is particularly

significant in the high SNR regime (Marshoud et al., 2016). This phenomenon can be explained by using CR-NOMA as an example. Recall that CR-NOMA first provides sufficient power to meet some users' QoS requirements. In the high SNR regime, the users' predefined QoS requirements can be easily met, and there will be ample power left to serve additional users, which yields the significant gain of NOMA over OMA. In VLC, it is typical that there is a strong line-of-sight path between the transceivers and the distance between the transmit light-emitting diode (LED) and that the receive photo detector (PD) is short, which means that VLC operates mainly in the high SNR regime and hence it is ideal for the application of NOMA. VLC is typically applied to a cell with a small coverage (Zhang X et al., 2017). Therefore, very few devices are connected within the small-size cell, which is helpful to reduce the complexity for the implementation of NOMA. Furthermore, VLC channels are significantly affected by the transmission angles of the transmit LEDs and the fields of view (FOVs) of the PD, which are new system parameters not presented in conventional radio frequency systems. By adjusting these system parameters, the channel conditions of the users can be dynamically controlled to facilitate the implementation of NOMA.

4.3 Terrestrial TV broadcasting and terrestrial-satellite communications

Terrestrial digital TV broadcasting is surprisingly becoming one of the first practical systems to which NOMA has been applied. Conventionally orthogonal multiplexing techniques, such as frequency division multiplexing and time division multiplexing, have been used for TV broadcasting, due to their low system complexity and affordable costs. However, the spectral efficiency of these orthogonal multiplexing techniques is low, and cannot be used to meet users' diverse QoS requirements. Recently, the Advanced Television Systems Committee (ATSC) has proposed a new type of terrestrial TV broadcasting, for which the corresponding physical layer protocol standard is known as ATSC 3.0 (Fay et al., 2016). In this new generation of digital TV standards, a new type of multiplexing based on the NOMA principle—layered division multiplexing (LDM)—has been used.

The key idea of LDM is very similar to power-domain NOMA, in that multiple broadcasting services are integrated on a single bandwidth resource block (Zhang L et al., 2016). Particularly, the simplest form of LDM integrates two layers, namely a core layer and an enhanced layer, at the same time and frequency. The signals from the two layers are intended to destinations with different receive capabilities. These signals are encoded with different types of error correction codes and then are superimposed in the same manner as power-domain NOMA. The benefit of LDM can be explained by the following example. Suppose that there are two types of TV broadcasting receivers. One type of receiver is static and has strong connections with the TV broadcasting station, and the other can be mobile receivers, such as pedestrians and vehicles. By using LDM, two types of broadcasting services, high-definition services and ultra-high-definition (UHD) streaming, such as 4K UHD and 8K UHD televisions, can be integrated on a single resource block and broadcast to the users. Similar to the receivers for power-domain NOMA, the users with weak channel conditions can at least decode the signals in the core layer, by treating the information in the enhanced layer as noise. The users with strong channel conditions can receive UHD video streaming by carrying out SIC and decoding the signals in the core layer before decoding UHD signals.

NOMA is also useful for terrestrial-satellite communications (Caus et al., 2016; Zhu et al., 2017). Conventional cellular networks can be viewed as a special case of terrestrial communications, which support a large system throughput but suffer from a lack of global coverage. However, satellite communications provide seamless global coverage, which motivates the joint design of terrestrial-satellite communications. For this new type of communications, researchers have shown that the application of NOMA enables a heterogenous network architecture, in which users are jointly served by a satellite and terrestrial base stations. Due to the large distances between the satellite and the users, the signals from these two communication systems need to be carefully structured for interference management purposes, where user clustering/grouping has a significant impact on the overall performance of the NOMA-assisted terrestrial-satellite communication system.

5 Future directions

5.1 NOMA with imperfect CSI

In NOMA, users are allowed to use the same bandwidth resource block, which means that strong co-channel interference exists in NOMA systems and is suppressed by using advanced signal processing algorithms at the transceivers, such as SIC, power allocation, and beamforming/precoding. Typically the implementation of these signal processing algorithms requires perfect CSI at the transmitter, although some NOMA schemes in Saito et al. (2013) and Ding et al. (2016a) do not make strong assumptions about CSI. For example, the simplest form of power-domain NOMA requires the base station to know only the order of users' channels, instead of the exact channel gains. Furthermore, in Ding et al. (2016a), the users' channel matrices are not required to be available at the base stations, and only scalar effective channel gains are needed.

In practice, the transmitter can have access to imperfect CSI only, and imperfect CSI can be generally divided into three types. One is CSI with channel estimation error due to the use of imperfect channel estimators. Note that a straightforward method for channel estimation in NOMA is to assign orthogonal pilots to users, in the same manner as OMA; however, a more spectrally efficient method is to superimpose the unknown data with predefined pilot signals, which reduces the system overhead to send the training information (Zhou et al., 2003). The second type of imperfect CSI is statistical CSI, where instantaneous CSI is not available but the statistical information about CSI is known by the base station. This type of imperfect CSI is motivated by the fact that it is difficult for the base station to perfectly know its connections to the users. Take a downlink case with high-mobility users as an example. An inordinate amount of system overhead needs to be consumed to carry out frequent channel estimation at the users, and the problem becomes more challenging when the CSI fed from the users back to the base station is outdated. As a result, asking the users to feed statistical information about CSI, such as large-scale path loss which varies slowly, back to the base station becomes preferable. In Yang et al. (2016c), NOMA was shown to be robust to these two types of imperfect CSI, compared with OMA.

The third type of imperfect CSI is limited feedback, which again avoids using too much system overhead for channel feedback (Xu et al., 2016; Ding et al., 2017c). Unlike the previous two types of imperfect CSI, limited feedback offers some degrees of freedom to realize a balanced tradeoff between system performance and complexity, which is the reason why this direction has attracted considerable attention. Take one-bit feedback as an example, which asks the base station to broadcast a threshold and each user to feed 1 (0) back if its channel gain is above (below) the threshold. Obviously the threshold is an important system parameter which should be carefully designed. An appropriate choice of the threshold ensures that the maximal multi-user diversity gain is achievable in NOMA systems, even with one-bit feedback. In practice, one-bit feedback is an extreme case, and multiple bits might be afforded for channel feedback, which means that the number of feedback bits is another system parameter to be optimized.

5.2 Combining NOMA with full duplexing

Full duplexing is another important technology to be used in 5G (Lee and Quek, 2017), and its key idea is to enable a communication node to transmit and receive at the same time. Intuitively full duplexing is more spectrally efficient than half duplexing, as a node can carry out the two functionalities, transmitting and receiving, by using a single resource block. While the performance gain of full duplexing over half duplexing is clear, most existing communication systems are still based on the half duplexing mode, since full duplexing suffers from strong loopback interference; i.e., the transmitted signals become cross-talk interference to the received signal. However, thanks to the recent advances in loopback interference cancellation techniques, full duplexing has attracted substantial attention during the past few years.

The most well known example for the combination of full duplexing and NOMA is the application of full duplexing to cooperative NOMA (Zhong and Zhang, 2016; Zhang L et al., 2017; Zhang Z et al., 2017a). Take cooperative NOMA without dedicated relays as an example. As discussed in Section 3.4, users with strong channel conditions are employed as relays to help those with poor channel conditions. While this cooperative NOMA can improve the reception reliability of the weak users, the overall

spectral efficiency of NOMA is reduced, since extra time slots are consumed for carrying out relay transmission. The application of full duplexing to cooperative NOMA means that strong users receive signals from their base station while helping weak users. As a result, no extra time slot is required for relay transmission, compared with non-cooperative NOMA, which demonstrates that full duplexing is particularly important for cooperative NOMA. Current full duplexing techniques cannot guarantee complete removal of the loopback interference, and many researchers are working on identifying the impact of the residual loopback interference on cooperative NOMA with full duplexing.

In addition to cooperative NOMA, full duplexing can be applied to other types of NOMA scenarios. For example, full duplexing can be used to realize simultaneous uplink and downlink transmission (Elbamby et al., 2017; Sun et al., 2017; Ding et al., 2018). In particular, a base station with full duplexing capability can broadcast the NOMA mixture to the users, while receiving the uplink signals from the users. If the half duplexing mode is used, double the bandwidth resources are needed, compared with the full duplexing case. Similar to cooperative NOMA, the residual loopback interference can potentially degrade the system performance; i.e., the reception reliability of the base station to decode the uplink signals is deteriorated by the transmitted downlink signals. However, unlike cooperative NOMA, the joint design of uplink and downlink results in another issue that uplink users can cause strong interference to downlink users. Because uplink and downlink transmissions happen at the same time, the performance of downlink users can be significantly degraded by nearby uplink users. However, in Ding et al. (2018), it was shown that significant performance gains over half duplexing NOMA and full duplexing OMA can still be realized, if uplink and downlink users are placed in different sectors of a cell. More sophisticated algorithms for uplink and downlink user clustering and scheduling are needed to avoid this co-channel interference, by dynamically using the users' channel conditions.

5.3 NOMA-assisted physical layer security

Similar to conventional multiple access techniques, security was not considered when the concept of NOMA was originally conceived. This is because in the current mobile systems, security is realized by using upper layer cryptographic methods. For example, in OFDMA systems, a user that is allocated to the first subcarrier is capable of decoding the bits sent on the other subcarriers, but the use of cryptographic methods ensures that this user cannot know the meaning of the decoded bits. However, initial studies have indicated that the use of NOMA transmission is helpful to the implementation of physical layer approaches to security, as illustrated below.

In the presence of external eavesdroppers, it has been shown that NOMA can improve the secrecy rates of the legitimate users in various communication scenarios compared with OMA (Zhang Y et al., 2016, 2017; Xu et al., 2017). Take the simple downlink power-domain NOMA scheme as an example, where the base station superimposes the users' signals and broadcasts the generated mixture. Compared with OMA cases, signals are not separately sent in NOMA, and the power allocation coefficients are tailored to the channel conditions of the NOMA users, which makes it more difficult for the eavesdroppers to intercept the signals sent to the legitimate receivers. However, there are two key challenges for secure NOMA transmission with external eavesdroppers. One is how to rigorously define the decoding rates at the eavesdroppers. A straightforward way is to assume that the eavesdroppers use the same SIC procedure as the legitimate users. However, there are alternatives for the eavesdropping strategies; e.g., a signal decoded during the last stage of SIC at legitimate users might be decoded first at the eavesdroppers, which can have a significant impact on the secrecy rates. The other is that different users' signals are protected in an unequal way, since the users' signals are transmitted with different power levels and hence the users experience different security performance. Advanced resource allocation algorithms are needed to meet users' different security requirements.

Compared with the scenario with external eavesdroppers, the scenario in which some NOMA users are potential eavesdroppers is more challenging. For example, how to avoid a strong NOMA user decoding a weak user's information is an important question, to which no answer has yet been found. However, for some special cases, NOMA transmission has been shown to be more secure than OMA, even if some legitimate users are potential eavesdrop-

pers. For example, in Ding et al. (2017e), multicasting and unicasting services are integrated together by the NOMA principle, where multicasting receivers are legitimate users in the system but may want to intercept the signal sent to the unicasting receiver. Without NOMA, unicasting and multicasting transmissions are separated, which means that the multicasting receivers can easily intercept the unicasting signals. By using the NOMA principle, the base station can use the multicasting signals as a type of jamming information to degrade the capabilities of the multicasting receivers for intercepting the unicasting signals. However, it is still not clear whether this idea can be extended to general NOMA scenarios beyond joint multicast-unicast transmission.

5.4 NOMA-assisted radio frequency and VLC-based energy harvesting

Radio frequency based energy harvesting (RFEH), also termed 'simultaneous wireless information and power transfer (SWIPT)', uses radio frequency signals for two purposes, namely energy harvesting and information delivery. RFEH is particularly important to energy constrained wireless networks, where communication nodes have limited energy supplies and have no access to conventional energy sources (Liu et al., 2013). The interaction between the two technologies, NOMA and RFEH, is bidirectional. On the one hand, NOMA can be very useful for efficient RFEH; e.g., a base station can superimpose two types of signals by using the NOMA principle, one for information transfer and the other for energy harvesting. On the other hand, RFEH is important to many NOMA communication scenarios (Liu et al., 2016). Take cooperative NOMA as an example, in which a strong user acts as a relay and helps a weak user. Relay transmission consumes the strong user's battery life to save the overall energy, and thus the strong user might not want to help the weak user. By applying RFEH, the strong user can harvest some energy from the radio frequency signals sent by the base station, which will be used to power the relay transmission. In this way, the strong user helps the weak user without reducing its own battery life, which provides an incentive for user cooperation. In addition to cooperative NOMA, RFEH has been shown to be useful for joint NOMA uplink and downlink transmissions (Diamantoulakis et al., 2016). RFEH couples the transmission power and users' channels, which means that the performance analysis and system optimization in NOMA-RFEH are more challenging than in conventional networks.

Using radio frequency signals is not the only way to charge nodes wirelessly, and recently the use of visible light as an alternative for RFEH has attracted attention (Pan et al., 2017a, 2017b). The main motivation to use visible light instead of radio frequency signals for energy harvesting is safety, as explained below:

1. At the transmitter side, there is a restriction on the maximum radio frequency transmission power, because of the public concern about electromagnetic pollution. Because of this power constraint and the path loss of radio frequency propagation, the amount of energy harvested at the receiver-based RFEH is quite limited.

2. At the receiver side, the receive power density is also strictly regulated, which means that the use of advanced smart antenna techniques to beam the energy to the desired location and hence the energy efficiency improvement may not be allowed.

Compared with the aforementioned difficulty of RFEH, visible light will cause fewer safety problems. In addition, compared with energy harvesting circuits using radio frequency signals, those using light for energy harvesting are much more mature and cheaper. Despite these advantages, VLC channels are not as well understood as radio frequency channels, which imposes many challenges on the design and analysis of the system using light for energy harvesting and information transfer.

6 Conclusions

In this study, we have provided a detailed survey of the impact of the emerging communication technique, NOMA, on future wireless networks. Particularly, how the NOMA principle affects the design of the next generation of multiple access techniques has been introduced, where different practical forms of NOMA have been described. Then the applications of NOMA to other advanced communication techniques, such as wireless caching, MIMO techniques, mmWave communications, and cooperative relaying, have been discussed. The impact of NOMA on communication systems beyond mobile communication networks has also been illustrated, through the examples of digital TV, satellite communications,

vehicular networks, and VLC. Finally, important directions for future research on NOMA, such as the implementation of NOMA with imperfect CSI, the applications of NOMA to physical layer security, energy harvesting, and full duplex transmission, have also been outlined.

References

Alkhateeb A, Nam YH, Zhang J, et al., 2016. Massive MIMO combining with switches. *IEEE Wirel Commun Lett*, 5(3):232-235. https://doi.org/10.1109/LWC.2016.2522963

Bastug E, Bennis M, Debbah M, 2014. Living on the edge: the role of proactive caching in 5G wireless networks. *IEEE Commun Mag*, 52(8):82-89. https://doi.org/10.1109/MCOM.2014.6871674

Boyd S, Vandenberghe L, 2004. Convex Optimization. Cambridge University Press, New York, USA.

Cai D, Fan P, Lei X, et al., 2016. Multi-dimensional SCMA codebook design based on constellation rotation and interleaving. IEEE 83rd Vehicular Technology Conf, p.1-5. https://doi.org/10.1109/VTCSpring.2016.7504356

Caus M, Vázquez MA, Pérez-Neira A, 2016. NOMA and interference limited satellite scenarios. 50th Asilomar Conf on Signals, Systems and Computers, p.497-501. https://doi.org/10.1109/ACSSC.2016.7869089

Chen S, Ren B, Gao Q, et al., 2017a. Pattern division multiple access (PDMA)—a novel non-orthogonal multiple access for 5G radio networks. *IEEE Trans Veh Technol*, 66(4):3185-3196. https://doi.org/10.1109/TVT.2016.2596438

Chen S, Hu J, Shi Y, et al., 2017b. Vehicle-to-everything (V2X) services supported by LTE-based systems and 5G. *IEEE Commun Stand Mag*, 1(2):70-76. https://doi.org/10.1109/MCOMSTD.2017.1700015

Chen Y, Wang L, Ai Y, et al., 2017. Performance analysis of NOMA-SM in vehicle-to-vehicle massive MIMO channels. *IEEE J Sel Areas Commun*, 35(12):2653-2666. https://doi.org/10.1109/JSAC.2017.2726006

Chen Z, Kountouris M, 2016. D2D caching vs. small cell caching: where to cache content in a wireless network? IEEE 17th Int Workshop on Signal Processing Advances in Wireless Communications, p.1-6. https://doi.org/10.1109/SPAWC.2016.7536874

Chen Z, Ding Z, Dai X, et al., 2016a. On the application of quasi-degradation to MISO-NOMA downlink. *IEEE Trans Signal Process*, 64(23):6174-6189. https://doi.org/10.1109/TSP.2016.2603971

Chen Z, Ding Z, 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

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 transmissions. *IEEE Trans Wirel Commun*, 15(5):3226-3237. https://doi.org/10.1109/TWC.2016.2518182

Cover TM, Thomas JA, 2006. Elements of Information Theory. John Wiley and Sons, New Jersey, USA. https://doi.org/10.1002/047174882X

Di B, Song L, Li Y, et al., 2017. Non-orthogonal multiple access for high-reliable and low-latency V2X communications in 5G systems. *IEEE J Sel Areas Commun*, 35(10):2383-2397. https://doi.org/10.1109/JSAC.2017.2726018

Diamantoulakis PD, Pappi KN, Ding Z, et al., 2016. Wireless-powered communications with non-orthogonal multiple access. *IEEE Trans Wirel Commun*, 15(12):8422-8436. https://doi.org/10.1109/TWC.2016.2614937

Ding Z, Poor HV, 2016. Design of massive-MIMO-NOMA with limited feedback. *IEEE Signal Process Lett*, 23(5):629-633. https://doi.org/10.1109/LSP.2016.2543025

Ding Z, Yang Z, Fan P, 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 Z, Peng M, Poor HV, 2015. Cooperative non-orthogonal multiple access in 5G systems. *IEEE Commun Lett*, 19(8):1462-1465. https://doi.org/10.1109/LCOMM.2015.2441064

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

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

Ding Z, Dai L, Poor HV, 2016d. MIMO-NOMA design for small packet transmission in the Internet of Things. *IEEE Access*, 4:1393-1405. https://doi.org/10.1109/ACCESS.2016.2551040

Ding Z, Dai H, Poor HV, 2016e. Relay selection for cooperative NOMA. *IEEE Wirel Commun Lett*, 5(4):416-419. https://doi.org/10.1109/LWC.2016.2574709

Ding Z, Liu Y, Choi J, et al., 2017a. Application of non-orthogonal multiple access in LTE and 5G networks. *IEEE Commun Mag*, 55(2):185-191. https://doi.org/10.1109/MCOM.2017.1500657CM

Ding Z, Fan P, Karagiannidis G, et al., 2017b. NOMA assisted wireless caching: strategies and performance analysis. https://arxiv.org/abs/1709.06951

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

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

Ding Z, Zhao Z, Peng M, et al., 2017e. 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 Z, Lei X, Karagiannidis GK, et al., 2017f. 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 Z, Fan P, Poor HV, 2018. On the coexistence between full-duplex and NOMA. *IEEE Wirel Commun Lett*, in press. https://doi.org/10.1109/LWC.2018.2811492

Elbamby MS, Bennis M, Saad W, et al., 2017. Resource optimization and power allocation in full duplex non-orthogonal multiple access (FD-NOMA) networks. *IEEE J Sel Areas Commun*, 35(12):2860-2873. https://doi.org/10.1109/JSAC.2017.2726218

Fay L, Michael L, Gómez-Barquero D, et al., 2016. An overview of the ATSC 3.0 physical layer specification. *IEEE Trans Broadcast*, 62(1):159-171. https://doi.org/10.1109/TBC.2015.2505417

Foschini GJ, Gans MJ, 1998. On limits of wireless communication in a fading environment when using multiple antennas. *Wirel Pers Commun*, 6(3):311-335. https://doi.org/10.1023/A:1008889222784

Gao X, Dai L, Sun Y, et al., 2017. Machine learning inspired energy-efficient hybrid precoding for mmWave massive MIMO systems. IEEE Int Conf on Communications, p.1-6. https://doi.org/10.1109/ICC.2017.7997065

Golrezaei N, Molisch AF, Dimakis AG, et al., 2013. Femtocaching and device-to-device collaboration: a new architecture for wireless video distribution. *IEEE Commun Mag*, 51(4):142-149. https://doi.org/10.1109/MCOM.2013.6495773

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

Heath RW, González-Prelcic N, Rangan S, et al., 2016. An overview of signal processing techniques for millimeter wave MIMO systems. *IEEE J Sel Topics Signal Process*, 10(3):436-453. https://doi.org/10.1109/JSTSP.2016.2523924

Ho IWH, Leung KK, Polak JW, 2011. Stochastic model and connectivity dynamics for VANETs in signalized road systems. *IEEE/ACM Trans Network*, 19(1):195-208. https://doi.org/10.1109/TNET.2010.2057257

Huawei Inc., 2015. 5G: a Techology Vision. http://www.huawei.com/en/about-huawei/publications/winwin-magazine/19/HW_329327

Kim JB, Lee IH, 2015. Non-orthogonal multiple access in coordinated direct and relay transmission. *IEEE Commun Lett*, 19(11):2037-2040. https://doi.org/10.1109/LCOMM.2015.2474856

Komine T, Nakagawa M, 2004. Fundamental analysis for visible-light communication system using LED lights. *IEEE Trans Consum Electron*, 50(1):100-107. https://doi.org/10.1109/TCE.2004.1277847

Kulkarni MN, Ghosh A, Andrews JG, 2016. A comparison of MIMO techniques in downlink millimeter wave cellular networks with hybrid beamforming. *IEEE Trans Commun*, 64(5):1952-1967. https://doi.org/10.1109/TCOMM.2016.2542825

Lee J, Quek TQS, 2017. Hybrid full-/half-duplex system analysis in heterogeneous wireless network. *IEEE Trans Wirel Commun*, 14(5):2883-2895. https://doi.org/10.1109/TWC.2015.2396066

Liu L, Zhang R, Chua KC, 2013. Wireless information transfer with opportunistic energy harvesting. *IEEE Trans Wirel Commun*, 12(1):288-300. https://doi.org/10.1109/TWC.2012.113012.120500

Liu Y, Ding Z, Elkashlan M, et al., 2016. Cooperative non-orthogonal multiple access with simultaneous wireless information and power transfer. *IEEE J Sel Areas Commun*, 34(4):938-953. https://doi.org/10.1109/JSAC.2016.2549378

Luo S, Teh KC, 2017. Adaptive transmission for cooperative NOMA system with buffer-aided relaying. *IEEE Commun Lett*, 21(4):937-940. https://doi.org/10.1109/LCOMM.2016.2647250

Lv L, Ni Q, Ding Z, 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

Maddah-Ali MA, Niesen U, 2014. Fundamental limits of caching. *IEEE Trans Inform Theory*, 60(5):2856-2867. https://doi.org/10.1109/TIT.2014.2306938

Marshoud H, Kapinas VM, Karagiannidis GK, et al., 2016. Non-orthogonal multiple access for visible light communications. *IEEE Photon Technol Lett*, 28(1):51-54. https://doi.org/10.1109/LPT.2015.2479600

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

Molina-Masegosa R, Gozalvez J, 2017. LTE-V for sidelink 5G V2X vehicular communications: a new 5G technology for short-range vehicle-to-everything communications. *IEEE Veh Technol Mag*, 12(4):30-39. https://doi.org/10.1109/MVT.2017.2752798

Nikopour H, Baligh H, 2013. Sparse code multiple access. IEEE 24$^{\text{th}}$ Int Symp on Personal Indoor and Mobile Radio Communications, p.332-336. https://doi.org/10.1109/PIMRC.2013.6666156

Nonaka N, Benjebbour A, Higuchi K, 2014. System-level throughput of NOMA using intra-beam superposition coding and SIC in MIMO downlink when channel estimation error exists. IEEE Int Conf on Communication Systems, p.202-206. https://doi.org/10.1109/ICCS.2014.7024794

NTT Docomo Inc., 2014. 5G Radio Access: Requirements, Concepts and Technologies.

NTT Docomo Inc., 2017. World's First Successful 5G Trial Using Smartphone-Sized NOMA Chipset-Embedded Device to Increase Spectral Efficiency. https://www.nttdocomo.co.jp/english/info/media_center/pr/2017/1102_02.html

Pan G, Ye J, Ding Z, 2017a. Secure hybrid VLC-RF systems with light energy harvesting. *IEEE Trans Commun*, 65(10):4348-4359. https://doi.org/10.1109/TCOMM.2017.2709314

Pan G, Ye J, Ding Z, 2017b. On secure VLC systems with spatially random terminals. *IEEE Commun Lett*, 21(3):492-495. https://doi.org/10.1109/LCOMM.2016.2643632

Proakis J, 2000. Digital Communications. McGraw-Hill, New York, USA.

Saito Y, Benjebbour A, Kishiyama Y, et al., 2013. System-level performance evaluation of downlink non-orthogonal multiple access (NOMA). IEEE 24$^{\text{th}}$ Int

Symp on Personal Indoor and Mobile Radio Communications, p.611-615. https://doi.org/10.1109/PIMRC.2013.6666209

Sun Y, Ng DWK, Ding Z, et al., 2017. Optimal joint power and subcarrier allocation for full-duplex multicarrier non-orthogonal multiple access systems. *IEEE Trans Commun*, 65(3):1077-1091. https://doi.org/10.1109/TCOMM.2017.2650992

Taherzadeh M, Nikopour H, Bayesteh A, et al., 2014. SCMA codebook design. IEEE 80th Vehicular Technology Conf, p.1-5. https://doi.org/10.1109/VTCFall.2014.6966170

techUK, 2015. 5G Innovation Opportunities—a Discussion Paper.

Verdú S, 1998. Multiuser Detection. Cambridge University Press, Cambridge, UK.

Wei Z, Yuan J, Ng D, et al., 2016. A survey of downlink non-orthogonal multiple access for 5G wireless communication networks. *ZTE Commun*, 14(4):17-25.

Wei Z, Dai L, Ng DWK, et al., 2017. Performance analysis of a hybrid downlink-uplink cooperative NOMA scheme. IEEE 85th Vehicular Technology Conf, p.1-7. https://doi.org/10.1109/VTCSpring.2017.8108407

Xu D, Ren P, Du Q, et al., 2017. Combat eavesdropping by full-duplex technology and signal transformation in non-orthogonal multiple access transmission. IEEE Int Conf on Communications, p.1-6. https://doi.org/10.1109/ICC.2017.7997115

Xu P, Ding Z, Dai X, et al., 2015. A new evaluation criterion for non-orthogonal multiple access in 5G software defined networks. *IEEE Access*, 3:1633-1639. https://doi.org/10.1109/ACCESS.2015.2480117

Xu P, Yuan Y, Ding Z, et al., 2016. On the outage performance of non-orthogonal multiple access with 1-bit feedback. *IEEE Trans Wirel Commun*, 15(10):6716-6730. https://doi.org/10.1109/TWC.2016.2587880

Xu X, Tao M, 2017. Modeling, analysis, and optimization of coded caching in small-cell networks. *IEEE Trans Commun*, 65(8):3415-3428. https://doi.org/10.1109/TCOMM.2017.2706726

Yakou K, Higuchi K, 2015. Downlink NOMA with SIC using unified user grouping for non-orthogonal user multiplexing and decoding order. Int Symp on Intelligent Signal Processing and Communication Systems, p.508-513. https://doi.org/10.1109/ISPACS.2015.7432825

Yang Z, Ding Z, Fan P, et al., 2016a. 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 Z, Cui J, Lei X, et al., 2016b. Impact of factor graph on average sum rate for uplink sparse code multiple access systems. *IEEE Access*, 4:6585-6590. https://doi.org/10.1109/ACCESS.2016.2614330

Yang Z, Ding Z, Fan P, et al., 2016c. On the performance of non-orthogonal multiple access systems with partial channel information. *IEEE Trans Commun*, 64(2):654-667. https://doi.org/10.1109/TCOMM.2015.2511078

Yang Z, Ding Z, Wu Y, et al., 2017. Novel relay selection strategies for cooperative NOMA. *IEEE Trans Veh Technol*, 66(11):10114-10123. https://doi.org/10.1109/TVT.2017.2752264

Yin L, Popoola WO, Wu X, et al., 2016. Performance evaluation of non-orthogonal multiple access in visible light communication. *IEEE Trans Commun*, 64(12):5162-5175. https://doi.org/10.1109/TCOMM.2016.2612195

Yu L, Fan P, Ma Z, et al., 2016. An optimized design of irregular SCMA codebook based on rotated angles and EXIT chart. IEEE 84th Vehicular Technology Conf, p.1-5. https://doi.org/10.1109/VTCFall.2016.7880904

Yu L, Fan P, Lei X, et al., 2017. BER analysis of SCMA systems with codebooks based on star-QAM signaling constellations. *IEEE Commun Lett*, 21(9):1925-1928. https://doi.org/10.1109/LCOMM.2017.2704090

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

Zhang D, Liu Y, Ding Z, et al., 2017. Performance analysis of non-regenerative massive-MIMO-NOMA relay systems for 5G. *IEEE Trans Commun*, 65(11):4777-4790. https://doi.org/10.1109/TCOMM.2017.2739728

Zhang L, Li W, Wu Y, et al., 2016. Layered-division-multiplexing: theory and practice. *IEEE Trans Broadcast*, 62(1):216-232. https://doi.org/10.1109/TBC.2015.2505408

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

Zhang Y, Wang HM, Zheng TX, et al., 2017. Energy-efficient transmission design in non-orthogonal multiple access. *IEEE Trans Veh Technol*, 66(3):2852-2857. https://doi.org/10.1109/TVT.2016.2578949

Zhang Z, Ma Z, Xiao M, et al., 2017a. Full-duplex device-to-device aided cooperative non-orthogonal multiple access. *IEEE Trans Veh Technol*, 66(5):4467-4471. https://doi.org/10.1109/TVT.2016.2600102

Zhang Z, Ma Z, Xiao Y, et al., 2017b. Non-orthogonal multiple access for cooperative multicast millimeter wave wireless networks. *IEEE J Sel Areas Commun*, 35(8):1794-1808. https://doi.org/10.1109/JSAC.2017.2710918

Zhong C, Zhang Z, 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 GT, Viberg M, McKelvey T, 2003. A first-order statistical method for channel estimation. *IEEE Signal Process Lett*, 10(3):57-60. https://doi.org/10.1109/LSP.2002.807864

Zhu X, Jiang C, Kuang L, et al., 2017. Non-orthogonal multiple access based integrated terrestrial-satellite networks. *IEEE J Sel Areas Commun*, 35(10):2253-2267. https://doi.org/10.1109/JSAC.2017.2724478