

Modulation and Multiple Access for 5G Networks

Yunlong Cai, *Senior Member, IEEE*, Zhijin Qin^{ID}, *Member, IEEE*, Fangyu Cui,
Geoffrey Ye Li, *Fellow, IEEE*, and Julie A. McCann

Abstract—Fifth generation (5G) wireless networks face various challenges in order to support large-scale heterogeneous traffic and users, therefore new modulation and multiple access (MA) schemes are being developed to meet the changing demands. As this research space is ever increasing, it becomes more important to analyze the various approaches, therefore, in this paper we present a comprehensive overview of the most promising modulation and MA schemes for 5G networks. Unlike other surveys of 5G networks, this paper focuses on multiplexing techniques, including modulation techniques in orthogonal MA (OMA) and various types of non-OMA (NOMA) techniques. Specifically, we first introduce different types of modulation schemes, potential for OMA, and compare their performance in terms of spectral efficiency, out-of-band leakage, and bit-error rate. We then pay close attention to various types of NOMA candidates, including power-domain NOMA, code-domain NOMA, and NOMA multiplexing in multiple domains. From this exploration, we can identify the opportunities and challenges that will have the most significant impacts on modulation and MA designs for 5G networks.

Index Terms—5G, modulation, non-orthogonal multiple access.

I. INTRODUCTION

IN RECENT years, fifth generation (5G) wireless networks have attracted extensive research interest. According to the 3rd generation partnership project (3GPP) [1], [2], 5G networks should support three major families of applications, including enhanced mobile broadband (eMBB) [1], [2]; massive machine type communications (mMTC) [1], [2]; and ultra-reliable and low-latency communications (URLLC) [1], [2]. On top of this, enhanced vehicle-to-everything (eV2X) communications are also considered as an important service that should be supported by 5G networks [1]. These scenarios require massive connectivity with high system throughput and improved spectral

efficiency (SE) and impose significant challenges to the design of general 5G networks. In order to meet these new requirements, new modulation and multiple access (MA) schemes are being explored.

Orthogonal frequency division multiplexing (OFDM) [3]–[5] has been adopted in fourth generation (4G) networks. With an appropriate cyclic prefix (CP), OFDM is able to combat the delay spread of wireless channels with simple detection methods, which makes it a popular solution for current broadband transmission. However, traditional OFDM is unable to meet many new demands required for 5G networks. For example, in the mMTC scenario [1], [2], sensor nodes usually transmit different types of data asynchronously in narrow bands while OFDM requires different users to be highly synchronized, otherwise there will be large interference among adjacent subbands.

To address the new challenges in 5G networks, various types of modulation have been proposed, such as filtering, pulse shaping, and precoding to reduce the out-of-band (OOB) leakage of OFDM signals. Filtering [6]–[9] is the most straightforward approach to reduce the OOB leakage and with a properly designed filter, the leakage over the stop-band can be greatly suppressed. Pulse shaping [10]–[13] can be regarded as a type of subcarrier-based filtering that reduces overlaps between subcarriers even inside the band of a single user, however, it usually has a long tail in time domain according to the Heisenberg-Gabor uncertainty principle [14]. Introducing precoding [15]–[18] to transmit data before OFDM modulation is also an effective approach to reduce leakage. In addition to the aforementioned approaches to reduce the leakage of OFDM signals, some new types of modulations have also been proposed specifically for 5G networks. For example, to deal with high Doppler spread in eV2X scenarios, transmit data can be modulated in the delay-Doppler domain [19]. The above modulations can be used with orthogonal multiple access (OMA) in 5G networks.

OMA is core to all previous and current wireless networks; time-division multiple access (TDMA) and frequency-division multiple access (FDMA) are used in the second generation (2G) systems, code-division multiple access (CDMA) in the third generation (3G) systems, and orthogonal frequency division multiple access (OFDMA) in the 4G systems. For these systems, resource blocks are orthogonally divided in time, frequency, or code domains, and therefore there is minimal interference among adjacent blocks and makes signal detection relatively simple. However, OMA can only support limited numbers of users due to limitations in the numbers of orthogonal resources blocks, which limits the SE and the capacity

Manuscript received February 16, 2017; revised June 11, 2017 and September 13, 2017; accepted October 23, 2017. Date of publication October 26, 2017; date of current version February 26, 2018. Y. Cai is funded by the National Natural Science Foundation of China under Grant 61471319. Part of the work of Z. Qin was done when she was with Department of Computing, Imperial College London, London SW7 2AZ, U.K. (*Corresponding author: Zhijin Qin.*)

Y. Cai and F. Cui are with the College of Information Science and Electronic Engineering, Zhejiang University, Hangzhou 310027, China (e-mail: ylcai@zju.edu.cn; cfy531@zju.edu.cn).

Z. Qin is with the School of Computing and Communications, Lancaster University, Lancaster LA1 4YW, U.K. (e-mail: zhijin.qin@lancaster.ac.uk).

G. Y. Li is with the School of Electrical and Computer Engineering, Georgia Institute of Technology, Atlanta, GA 404385 USA (e-mail: liye@ece.gatech.edu).

J. A. McCann is with the Department of Computing, Imperial College London, London SW7 2AZ, U.K. (e-mail: jamm@imperial.ac.uk).

Digital Object Identifier 10.1109/COMST.2017.2766698

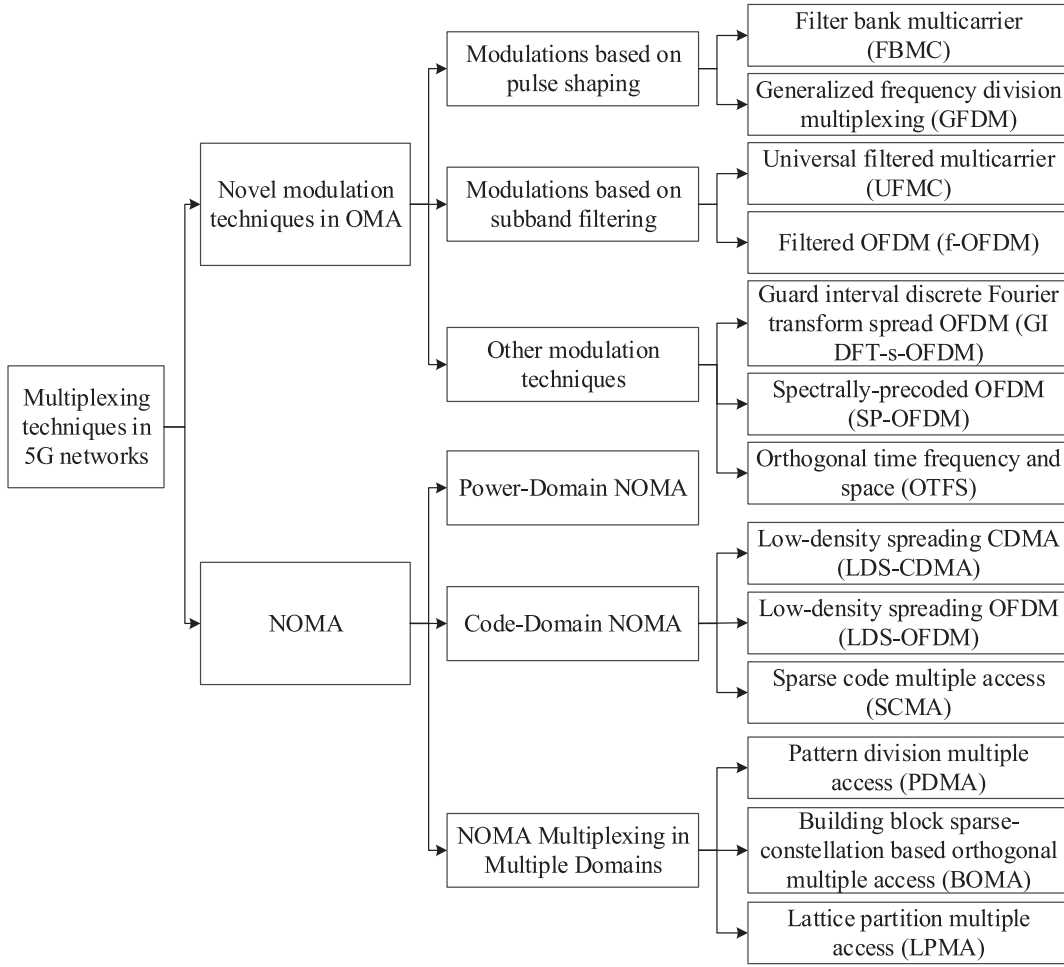


Fig. 1. A classification of the novel multiplexing techniques in 5G networks.

of current networks. To support a massive number of and dramatically different classes of users and applications in 5G networks, various NOMA schemes have been proposed.

As an alternative to OMA, NOMA introduces a new dimension by performing multiplexing within one of the classic time/frequency/code domains. In other words, NOMA can be regarded as an “add-on”, which has the potential to be harmoniously integrated with existing MA techniques. The core of NOMA is to utilize power and/or code domains in multiplexing to support more users in the same resource block. There are three major types of NOMA: power-domain NOMA, code-domain NOMA, and NOMA multiplexing in multiple domains. With NOMA, the limited spectrum resources can be fully utilized to support more users, therefore the capacity of 5G networks can be improved significantly even though extra interference and additional complexity will be introduced at the receiver.

This article, rather than discussing the whole structure of 5G networks [20], provides a comprehensive overview of multiplexing techniques, including modulation techniques in OMA and various types of NOMA schemes. To the best of our knowledge, existing surveys in this field mainly focus on either modulation or NOMA [21]–[25]. However, to address the various challenges in 5G networks, we can either develop novel

modulation techniques to reduce multiple user interference in OMA or directly use NOMA. Even if some surveys include both OMA and NOMA [26], [27], our survey includes a more complete introduction of different OMA and NOMA schemes, with a special emphasis on newly developed schemes. In this article, as shown in Fig. 1, we not only introduce OMA modulation and NOMA schemes, but also classify different schemes according to their basic principles. Additionally, we also compare different techniques in terms of performance and characteristics. Moreover, some open issues are also identified and discussed to inspire future research on modulation and MA in 5G networks. We hope that this article could help readers to build up a good knowledge of the promising multiplexing techniques in 5G networks.

The rest of this article is organized as follows. In Section II, novel modulation candidates for OMA in 5G networks are introduced and compared. In Section III, various NOMA schemes are discussed. Section IV concludes this article.

II. NOVEL MODULATION FOR OMA

In this section, we will discuss new modulation techniques for 5G networks. Since OFDM is widely used in current wireless systems and standards, many potential modulation

schemes for 5G networks are delivered from OFDM for backward compatibility reasons. Therefore, we will first introduce traditional OFDM.

A. Traditional OFDM

Denote d_k , for $k = 0, 1, \dots, N-1$, to be the transmit complex symbols. Then the baseband OFDM signal can be expressed as

$$s(t) = \sum_{k=0}^{N-1} d_k e^{j2\pi f_k t} \quad (1)$$

for $0 \leq t \leq T_s$, where $f_k = k\Delta f$, Δf is the subcarrier bandwidth and T_s is the symbol duration. To ensure that transmit symbols can be recovered without distortion, $\Delta f \cdot T_s = 1$, which is also called the orthogonal condition. It can be easily shown that

$$d_k = \frac{1}{T_s} \int_0^{T_s} s(t) e^{-j2\pi f_k t} dt \quad (2)$$

if the orthogonal condition holds.

Denote $s(n\Delta t)$ to be the sampled version of $s(t)$, where $\Delta t = \frac{T_s}{N}$. It can be easily seen [4] that $\{s(0), s(\Delta t), \dots, s((N-1)\Delta t)\}$ is the inverse discrete Fourier transform (IDFT) of $\{d_0, d_1, \dots, d_{N-1}\}$, which can be implemented by fast Fourier transform (FFT) and significantly simplifies OFDM modulation and demodulation.

To address the delay spread of wireless channels, a CP is usually used in OFDM. If the length of the CP is larger than the delay span (the duration between the first and the last taps/paths of a channel), then the demodulated OFDM signal can be expressed as

$$\hat{d}_k = H_k d_k + n_k, \quad (3)$$

where H_k is the frequency response of the wireless channel at $f_k = k\Delta f$ and n_k is the impact of additive channel noise. Therefore, the channel distortion becomes a multiplication of channel frequency response in OFDM systems while it is convolution in single-carrier systems, which makes the detection of OFDM signal much easier.

From the above discussion, OFDM can effectively deal with the delay spread of broadband wireless channels and FFT can be used to significantly simplify its complexity, therefore it has been widely used in the current wireless communication systems and standards. However, as we can see from (1), the OFDM signal is time-limited. Therefore, its OOB leakage is pretty high, especially when users are asynchronized as typical of 5G networks. To address this issue, a guard band is usually inserted between the signals of two adjacent users in the frequency domain in addition to a CP or a guard interval (GI) in the time domain, which reduces the SE of OFDM. This is even more severe for the users using a narrow frequency band.

5G networks have to support not only a massive number of users but also dramatically different types of users that have different demands. Traditional OFDM can no longer satisfy these requirements, and therefore novel modulation techniques with much lower OOB leakage are required. The new modulation techniques for 5G networks currently need to consider

backward compatibility with traditional OFDM systems but should also have the following key features to address the new challenges.

- 1) High SE: New modulation techniques should be able to mitigate OOB leakage among adjacent users so that the system SE can be improved significantly by reducing the guard band/time resources.
- 2) Loose synchronization requirements: Massive number of users are expected to be supported, especially for the Internet of Things (IoT), which makes synchronization difficult. Therefore, new modulation techniques are expected to accept asynchronous scenarios.
- 3) Flexibility: The modulation parameters (e.g., subcarrier width and symbol period) for each user should be configured independently and flexibly to support users with different data rate requirements.

There are various types of modulation techniques for OMA, which are based on different principles, including pulse shaping, subband filtering, precoding, GI shortening, and mapping symbols in the delay-Doppler domain. In the following, we will introduce these promising modulation techniques.

B. Modulations Based on Pulse Shaping

Pulse shaping, which is also regarded as subcarrier-based filtering, can effectively reduce OOB leakage. It is well known that the time and frequency widths of the pulses cannot be reduced at the same time. Therefore, the waveforms based on pulse shaping is usually non-orthogonal in both time and frequency domains to maintain high SE. Compared with traditional OFDM, the transceiver structure supporting pulse shaped modulation is more complex. Here, we introduce two typical modulations based on pulse shaping, i.e., filter bank multicarrier (FBMC) [10], [11] and generalized frequency division multiplexing (GFDM) [12], [13].

1) *FBMC*: As shown in Fig. 2, FBMC [10], [11] consists of IDFT and DFT, synthesis and analysis polyphase filter banks. The prototype filter in FBMC performs the pulse shaping. There are two types of typical pulses: the pulse based on the isotropic orthogonal transform algorithm (IOTA) [28] and the pulse adopted in the PHYDYAS project [29]. The length of the pulse in the time domain is determined by the required performance and is usually several times the length of the symbol period. The bandwidth of the pulse, which is different from the pulse in the traditional OFDM that has a long tail, is limited within a few subbands. To achieve the best SE, offset quadrature amplitude modulation (OQAM) is usually applied to make FBMC real-domain orthogonal in time and frequency domains [10]. Therefore, the transmit signal over $M/2$ consecutive block periods¹ can be expressed as

$$s(n) = \sum_{k=0}^{K-1} \sum_{m=0}^{M-1} d_{k,m} \theta_{k,m} g(n - mK/2) e^{j\frac{2\pi kn}{K}}, \quad (4)$$

where K and M are the numbers of subcarriers and symbols, respectively, $d_{k,m}$ is the transmit symbol at subcarrier

¹Block period corresponds to OFDM symbol period in the traditional OFDM systems.

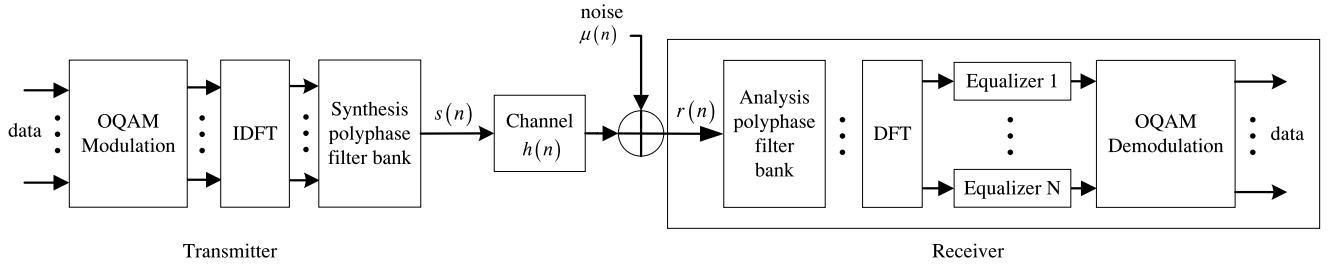


Fig. 2. Filter bank multicarrier (FBMC).

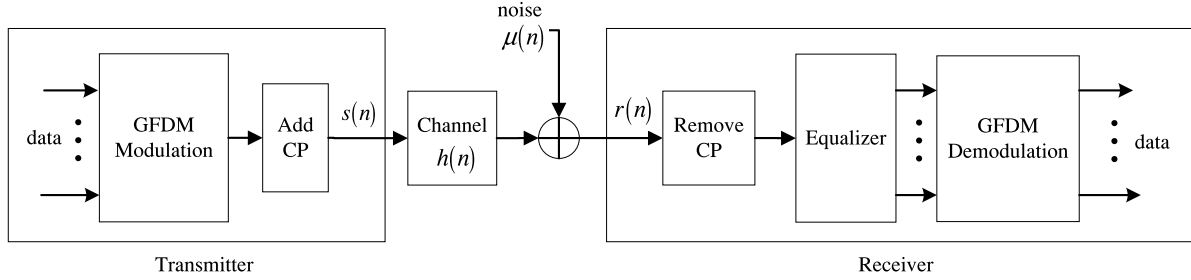


Fig. 3. Generalized frequency division multiplexing (GFDM).

k and symbol m , and $g(n)$ is the prototype filter coefficient at the n -th time-domain sample. It is worth noting that the transmit symbols here refer to the pulse amplitude modulation (PAM) symbols that are derived from the staggering of quadrature amplitude modulation (QAM) symbols. Thus the interval between two adjacent blocks is only half of the block period due to the offset in OQAM. The parameter, $\theta_{k,m}$ in (4), is defined as

$$\theta_{k,m} = \begin{cases} \pm 1, & \text{if } m+k \text{ is even,} \\ \pm j, & \text{if } m+k \text{ is odd,} \end{cases} \quad (5)$$

which is used to form the OQAM structure. With a properly designed prototype filter such as IOTA and the OQAM structure, the interference from the nearby overlapped symbols caused by a matched filter (MF) receiver becomes pure imaginary, which can be easily cancelled.

2) *GFDM*: Fig. 3 demonstrates the block diagram of GFDM. OFDM and single-carrier frequency division multiplexing (SC-FDM) can be regarded as two special cases of GFDM [13]. The unique feature of GFDM is to use circular shifted filters, rather than linear filters that are used in FBMC, to perform pulse shaping. By carefully choosing the circular filter, the out-of-block leakage can be reduced even if the orthogonality is completely given up. We can flexibly adjust M frequency samples and K time samples for a GFDM block according to the application environment. The transmit signal for each GFDM block can be expressed as

$$s(n) = \sum_{k=0}^{K-1} \sum_{m=0}^{M-1} d_{k,m} g_{k,m}(n), \quad (6)$$

for $0 \leq n \leq KM - 1$, where $d_{k,m}$ is the transmit symbol on subcarrier k at subsymbol m and $g_{k,m}(n)$ is the circular time and frequency shifted version of the prototype pulse shaping filter. In (6),

$$g_{k,m}(n) = g((n - mK)_{KM}) e^{\frac{j2\pi kn}{K}}, \quad (7)$$

where $(\cdot)_{KM}$ denotes the KM modulo operation and $g(n)$ is the prototype pulse shaping filter. Similar to the traditional OFDM, the modulation process and demodulation process can be expressed by matrix operations. The IDFT and DFT matrices in the traditional OFDM are substituted by some specific matrices corresponding to the modulation and demodulation for GFDM. But, the transceiver structure of GFDM is significantly different from the traditional OFDM.

Besides FBMC and GFDM, other modulations based on pulse shaping, such as pulse-shaped OFDM [30] and QAM-FBMC [31], have also been proposed for 5G networks. Generally, modulations based on pulse shaping try to restrict transmit signals within a narrow bandwidth and thus mitigate the OOB leakage so that they can work in asynchronous scenarios with a narrow guard band. FBMC also uses OQAM to achieve real-domain orthogonality, which saves the cost of the GI and interference cancellation. In addition, the circular shifted filters in GFDM avoid the long tail of the linear filters in the time domain, which makes GFDM fit for sporadic transmission. Furthermore, GFDM is easily compatible to multiple-input-multiple-output (MIMO) technologies [13].

C. Modulations Based on Subband Filtering

Subband filtering is another technique to reduce the OOB leakage. Universal filtered multicarrier (UFMC) [6], [7] and filtered OFDM (f-OFDM) [8], [9] are two typical modulations based on subband filtering, which will be introduced next.

1) *UFMC*: Fig. 4 shows the transmitter and the receiver structures of UFMC [6], [7]. In UFMC, the subbands are with equal size, and each filter is a shifted version of the same prototype filter. OFDM is applied within a subband for this modulation as shown in the figure. Since the bandwidth of the filter in UFMC is much wider than that of the modulations based on the pulse shaping, the length in time domain is much shorter. Therefore, interference caused by the tail of the

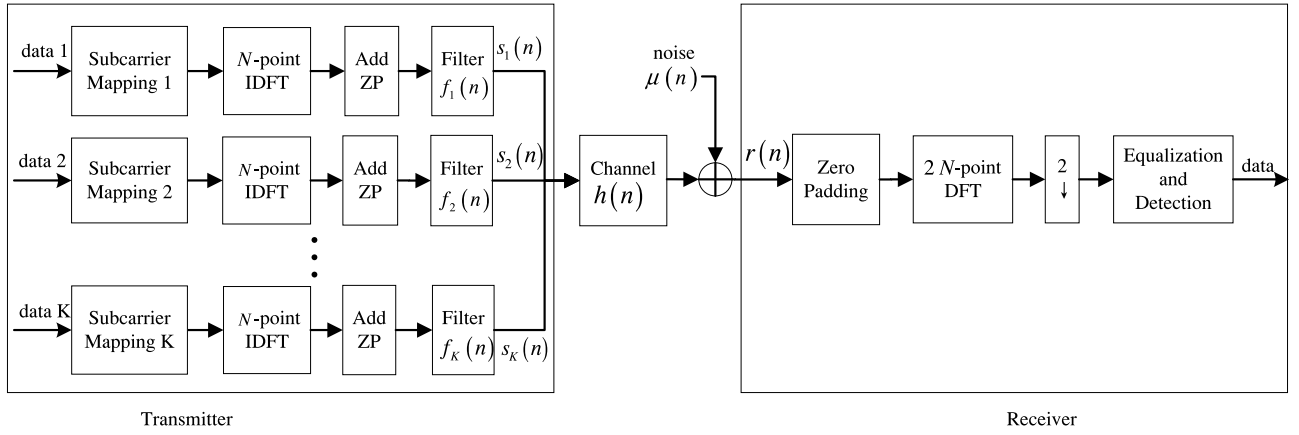


Fig. 4. Universal filtered multicarrier (UFMC).

filter can be easily eliminated by adopting a zero-padding (ZP) prefix with a reasonable length. Assuming that N subcarriers are divided into K subbands, each with $L = N/K$ consecutive subcarriers, the transmit signal in UFMC can be expressed as

$$s(n) = \sum_{k=0}^{K-1} s_k(n) * f_k(n), \quad (8)$$

where $f_k(n)$ is the filter coefficient of subband k , and $s_k(n)$ is the OFDM modulated signal over subband k that can be expressed as

$$s_k(n) = \sum_{m=0}^{M-1} s_{k,m}(n - m(N + N_g)) = \begin{cases} s_{k,m}(n - m(N + N_g)), & m(N + N_g) \leq n \leq m(N + N_g) + N - 1 \\ 0, & \text{otherwise} \end{cases} \quad (9)$$

with N_g denoting the length of the ZP, M denoting the number of symbol blocks and $s_{k,m}(n)$ denoting the signal at subcarrier k and symbol m . In (9), $s_{k,m}(n)$ can be expressed as

$$s_{k,m}(n) = \sum_{l=(k-1)L}^{kL-1} d_{l,m} e^{j \frac{2\pi l n}{N}}, \quad 0 \leq n \leq N - 1, \quad (10)$$

where $d_{l,m}$ denotes the l -th transmit symbol at the m -th symbol block.

At the receiver, the signal at each symbol interval is with the length of $N + N_g$ and is zero-padded to have a length of $2N$ so that a $2N$ -point FFT can be performed. Please note that only the even subcarriers are considered for signal detection after the $2N$ -point FFT.

2) *f-OFDM*: f-OFDM has a similar transmitter structure as UFMC [8], [9]. The main difference is that f-OFDM employs a CP and usually allows residual inter-symbol interference (ISI) [8]. Therefore, at the receiver, the MF is applied instead of the ZP and decimation. Besides, downsampling can be applied before the DFT operation, which can reduce complexity significantly since the CP can mitigate most of interference caused by the tail of the filter; the residual interference is with much lower power and can be treated as noise [9]. Thus,

the filter in f-OFDM can be longer than that in UFMC and has better attenuation outside the band. With the aid of effective channel coding, the performance degradation caused by residual interference in f-OFDM can be negligible. Another difference from UFMC is that the subcarrier spacing and the CP length do not have to be the same for different users in f-OFDM.

The most widely used filter in f-OFDM is the soft-truncated sinc filter [8], which can be easily used in various applications with different parameters. Therefore, f-OFDM is very flexible in the frequency multiplexing.

Besides UFMC and f-OFDM, other modulations based on subband filtering have also been proposed. For example, resource block f-OFDM (RB-f-OFDM) [32] utilizes filters based on resource block instead of the whole band of users in f-OFDM. In general, modulations based on subband filtering can effectively reduce OOB leakage and achieve better performance in comparison with the traditional OFDM.

D. Other Modulation Techniques

Apart from pulse shaping and subband filtering, there are also some other techniques to suppress the OOB leakage and meet the requirements of 5G networks. In the following, we mainly introduce three other modulations, including GI discrete Fourier transform spread OFDM (GI DFT-s-OFDM) [33], spectrally-precoded OFDM (SP-OFDM) [15], and orthogonal time frequency and space (OTFS) [19].

1) *GI DFT-s-OFDM*: In GI DFT-s-OFDM [33], the known sequence is used as the GI instead of a CP. Several types of the known sequences, such as the zero sequence [34] and a well-designed unique word [35], can be used. By a fixed known sequence with constant amplitude in GI DFT-s-OFDM, the peak-to-average power ratio (PAPR) of the modulated signal can be reduced. Moreover, the known sequence can also be utilized to estimate the parameters, such as the carrier frequency offset (CFO) in the synchronization process. By utilizing a proper sequence as the GI, the discontinuity between the adjacent time blocks in the traditional OFDM/DFT-s-OFDM can be avoided. As a result, the OOB leakage is reduced.

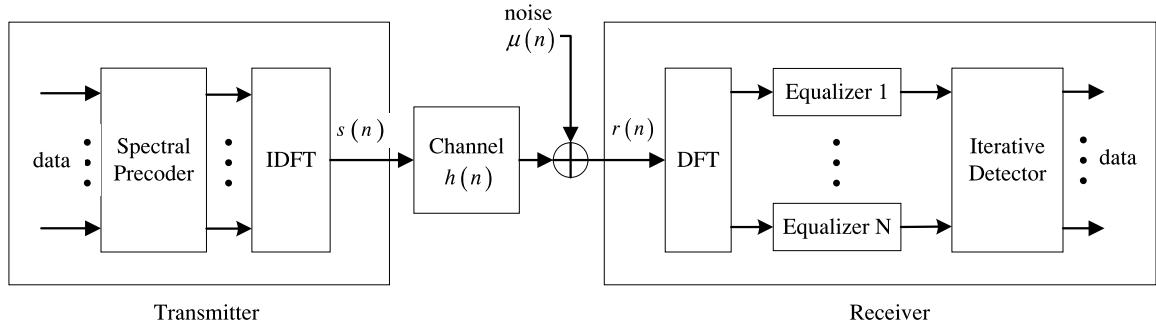


Fig. 5. Spectrally-precoded OFDM (SP-OFDM).

For GI DFT-s-OFDM, the overall length of the GI and useful signal for different users is same. Thus, the DFT windows for different users at the receiver can still be aligned even if the lengths of the GIs are different. Therefore, the mutual interference due to asynchronization of users can be mitigated [33].

2) *SP-OFDM*: Fig. 5 shows the diagram of SP-OFDM [15]. From the figure, it consists IDFT and DFT, spectral precoder, and iterative detector. Generally, the data symbols mapped on subcarriers are precoded by a rank-deficient matrix in order to project the signal into a properly selected lower dimensional subspace so that the precoded signal can be high-order continuous, and results in much lower leakage compared with the traditional OFDM [16], [17]. Even if precoded by a rank-deficient matrix can reduce the capacity of the channel, the OOB leakage of the OFDM signals can be significantly suppressed at the cost of only few reduced dimensions.

Compared to the modulations based on filtering, SP-OFDM has the following three advantages:

- The ISI caused by the tail of the filters can be removed without filtering. Therefore, the CP applied to combat the multipath of the wireless channels can be shorter, and SE is improved.
- When fragmented bands are used, SP-OFDM can easily notch specific well chosen frequencies without requiring multiple narrow subband filters [18].
- Furthermore, precoding and filtering can be combined to further improve the performance.

3) *OTFS*: The structure of OTFS is similar to SP-OFDM, as can be seen in Fig. 5. The main difference is that the spectral precoder and the iterative detector are substituted by the two-dimensional (2D) symplectic Fourier transform and the corresponding inverse transform modules. OTFS maps the symbols in the delay-Doppler domain [19]. Through a 2D symplectic Fourier transform, the corresponding data in the time-frequency domain can be calculated. Then, the calculated data can be transmitted via a time-frequency-domain modulation method as in OFDM. Since the 2D symplectic Fourier transform is relatively independent of the time-frequency-domain modulation method, pulse shaping and subband filtering can also be applied together to further reduce the leakage in OTFS.

When a mobile is with a high speed, the channel experiences fast fading. Channel parameters need to be estimated

and tracked very often therefore, which significantly increases resource costs. Moreover, most of the modulations are designed assuming that channels are constant within a symbol block. With a high mobility speeds, extra interference is introduced, which degrades the performance. However, in the delay-Doppler domain, the high Doppler channel can be expressed in a stable model, which saves the cost of tracking the time-varying fading and improves performance therefore.

OTFS can be also applied to estimate channel state information (CSI) of different antennas in MIMO systems [19]. Generally, the delay and Doppler dispersions are still relatively small compared to the system scale. In this case, the channel can be expressed in a compact and stable form in the delay-Doppler domain. As a result, the spread of the pilots caused by the channel are local, which enables to estimate the CSI of different antennas in MIMO systems by different pilots within a small area of the delay-Doppler plane.

In addition, a number of modulations based on other techniques have been also proposed, such as windowed OFDM (W-OFDM) [36], which utilizes windowing to deal with the discontinuity between adjacent OFDM symbols.

E. Performance Comparison of Different Modulations

We compare the power spectral density (PSD) and bit-error rate (BER) of different modulations.

1) *PSD*: Suppressing the OOB leakage is a key purpose for most of the modulation candidates for 5G networks. The PSDs of the some modulations are shown in Fig. 6. From the figure, all modulations achieve much lower leakage compared to the traditional OFDM. Among them, UPMC applies subband filtering and also has low leakage, and FBMC and f-OFDM have the lowest leakage. GFDM, GI DFT-s-OFDM, and SP-OFDM, although do not reduce the leakage as much as FBMC and f-OFDM, can still achieve much better performance than the traditional OFDM.

2) *BER*: In order to reduce the OOB leakage, many modulations utilize techniques, such as pulse shaping and subband filtering, which may introduce ISI and ICI. Hence, the BER performance of different modulations is compared here.

Fig. 7 shows the BER performance versus signal-to-noise ratio (SNR) when the Doppler spread $f_d = 0$ and 300 Hz. From Fig. 7 (a), the traditional OFDM has the best performance when the Doppler spread is zero ($f_d = 0$) since the ISI caused by the multipath has been completely canceled by

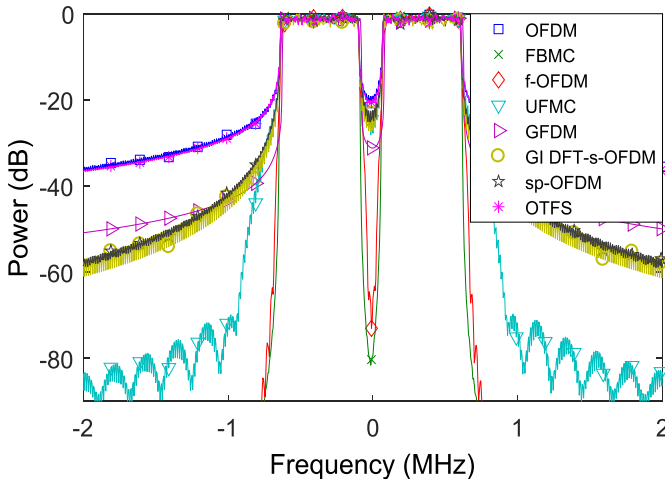


Fig. 6. PSDs of different modulations, where the bandwidth is 15.36 MHz, the DFT size is 1024, each of the two users occupies 36 subcarriers, and the guard band contains 12 subcarriers.

the CP. Since the bandwidth of each subcarrier is small enough to make the corresponding channel approximately flat, the ISI introduced by pulse shaping in FBMC is nearly pure imaginary. Therefore, FBMC is approximately orthogonal in the real domain and achieves good BER performance. The performance of UPMC, GFDM, and SP-OFDM is similar to that of FBMC, which is degraded slightly due to noise enhancement and low-projection precoding. However, f-OFDM introduces extra ISI that cannot be completely canceled, and as a result, it has slightly worse performance, especially in the high SNR region. GI DFT-s-OFDM and OTFS, which are different from the modulation schemes that directly map the symbols on subcarriers, apply spreading before mapping so that their performance does not approach that of OFDM. Since the fast-fading channel is difficult to be estimated and tracked accurately, the performance of the most modulation schemes degrades significantly as we can see from Fig. 7 (b). While OTFS can still achieve good performance due to its specific channel estimation method. Moreover, its performance in the high-mobility scenario is even better than that in the zero Doppler shift scenario because of Doppler diversity.

F. Future Work

In this section, modulation techniques for 5G networks are discussed. These techniques can be used with OMA to effectively deal with the OOB leakage in 5G networks. However, there are still many open issues in this area.

- 1) A potential application of f-OFDM is IoT. In this scenario, the subbands are narrow and therefore, interference caused by a short CP can significantly degrade the performance and should be considered in the detection. To improve the detection performance, additional processing, such as filtering or successive interference cancellation (SIC), is needed. Residual ISI cancellation (RISIC) [37] could be helpful.

- 2) The existing designs for subband filtering, such as Dolph-Chebyshev filter in UPMC and soft-truncated filter in f-OFDM, are with fixed length. However, different users and different application scenarios will have different requirements on the leakage levels, filter lengths, etc. Since the time and frequency dispersions are dual variables that cannot be reduced at the same time, how to design an efficient prototype filter to balance them according to application scenarios becomes interesting.
- 3) Similar to the traditional OFDM, multi-carrier based new modulation candidates, such as FBMC and UPMC also have a large PAPR. In order to improve the efficiency of the power amplifier, the PAPR should be reduced. The traditional PAPR reduction methods [38] applied in the traditional OFDM usually introduce distortions that degrade the performance. Therefore, how to properly extend the PAPR reduction methods in the traditional OFDM to the new modulations is an interesting and meaningful issue.

III. NOMA

In order to support higher throughput and massive and heterogeneous connectivity for 5G networks, we can adopt novel modulations discussed in Section II for OMA, or directly use NOMA with effective interference mitigation and signal detection methods. The key features of NOMA can be summarized as follows:

- 1) Improved SE: NOMA exhibits a high SE, which is attributed to the fact that it allows each resource block (e.g., time/frequency/code) to be exploited by multiple users.
- 2) Ultra high connectivity: With the capability to support multiple users within one resource block, NOMA can potentially support massive connectivity for billions of smart devices. This feature is quite essential for IoT scenarios with users that only require very low data rates but with massive number of users.
- 3) Relaxed channel feedback: In NOMA, perfect uplink CSI is not required at the base station (BS). Instead, only the received signal strength needs to be included in the channel feedback.
- 4) Low transmission latency: In the uplink of NOMA, there is no need to schedule requests from users to the BS, which is normally required in OMA schemes. As a result, a grant-free uplink transmission can be established in NOMA, which reduces the transmission latency drastically.

Existing NOMA schemes can be classified into three categories: power-domain NOMA, code-domain NOMA, and NOMA multiplexing in multiple domains. We will introduce them subsequently with emphasis on power-domain NOMA.

A. Power-Domain NOMA

Power-domain NOMA is considered as a promising MA scheme for 5G networks [39]. Specifically, a downlink

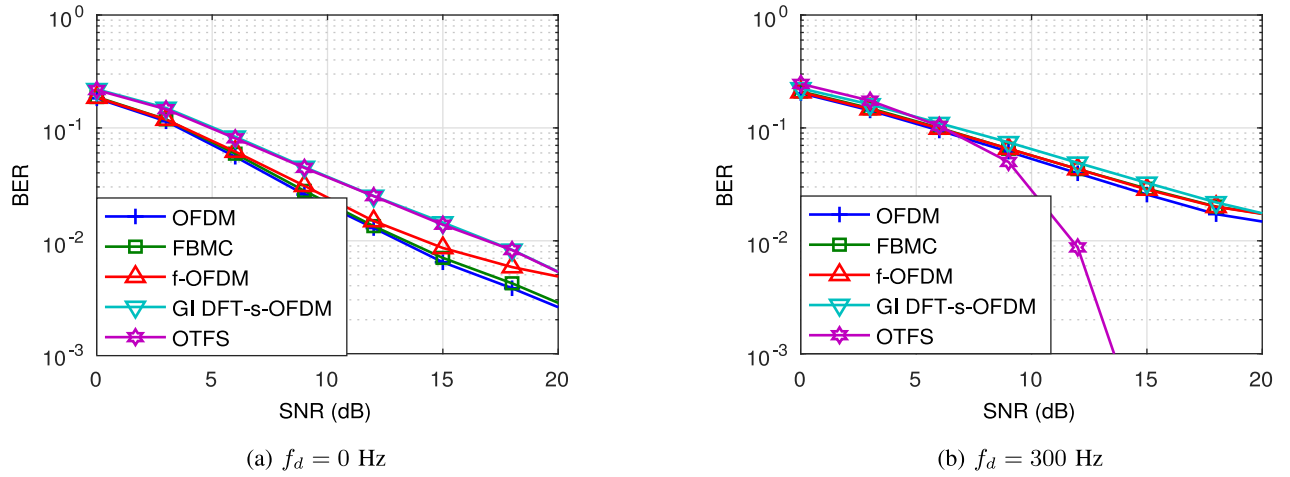


Fig. 7. BER versus SNR for different modulations, where the bandwidth is 15.36 MHz, the DFT size is 1024, each of the two users occupies 36 subcarriers, and the guard band contains 12 subcarriers for ITU Vehicular A channel model.

version of NOMA, named multiuser superposition transmission (MUST), has been proposed for the 3GPP long-term evolution advanced (3GPP-LTE-A) networks [40]. It has been shown that system capacity and user experiences can be improved by NOMA. More recently, a new work item (WI) outlining downlink multiuser superposition transmission for LTE has been approved by 3GPP LTE Release 14 [41], which aims to identify the necessary techniques to enable LTE to support the downlink intra-cell multiuser superposition transmission. In the following, we will expand upon the basic principles of various power-domain NOMA related techniques, including power allocation in NOMA, multiple antenna based NOMA, and cooperative NOMA.

1) *Basic Power-Domain NOMA*: Power-domain NOMA supports multiple users within the same time/frequency/code resource block by distinguishing them with different power levels. Unlike multiuser detection in CDMA or MIMO systems that have multiple observations at the receiver [42]–[46], power-domain NOMA usually only has one observation. Specifically, in the uplink transmission of NOMA, the signal received at the BS can be expressed as

$$Y = \sum_{a=1}^A h_a \sqrt{p_a} x_a + n, \quad (11)$$

where p_a and x_a are the transmit power and transmit symbols from the a th user, respectively, n refers to AWGN with variance σ^2 , and the number of users sharing the same resource block is A . The transmit power p_a for each individual user is carefully adjusted to facilitate SIC at the receiver, that is, to make sure users with stronger powers to be detected with high accuracy. At the receiver (the BS), with invoking of SIC, the user with the best CSI is decoded first. Then the corresponding signal component is removed from the received signal. The SIC receiver works in the descending order of the signal strengths. It should be noted that the transmit power levels of different NOMA users are usually different as users experience different channel conditions. If the first detected symbols are all correct, the received signal-to-interference-plus-noise ratio

(SINR) of the a th NOMA user can be given by

$$\text{SINR}_a = \frac{p_a |h_a|^2}{\sum_{b=a+1}^A p_b |h_b|^2 + \sigma^2}. \quad (12)$$

The downlink transmission of NOMA for the two-user case is shown in Fig. 8 where the users sharing the same resource block are distinguished by different power levels with a total power constraint. Specifically, the BS sends a superimposed signal containing the two signals for the two users. Different from conventional power allocation strategies, such as water filling, NOMA allocates less power for the users with better downlink CSI, to guarantee overall fairness and to utilize diversity in time/frequency/code domains. SIC is used for signal detection at the receiver. The user with higher transmit power, that is, the one with smaller downlink channel gain, is first decoded while treating another user's signal as noise. Once the signal corresponding to the user with higher transmit power is detected and decoded, its signal component will be subtracted from the received signal to facilitate the detection of subsequent users. It should be noted that the first detected user suffers from the highest inter-user interference and also the detection error in the first user will pass to the other users, which is why we have to allocate sufficient power to the first user to be detected. The extension of NOMA from two to multiple user cases is straightforward, and more related work on power allocation in NOMA is discussed in the following subsection.

Fig. 9 compares NOMA and OMA where two users are served by the same BS when NOMA is adopted. From the figure, the NOMA scheme achieves a lower outage probability. However, by adopting NOMA, a more complex transmitter and receiver are required to mitigate the interference. Furthermore, power-domain NOMA usually works well when only two or a few users share the same resource block. As the number of users multiplexing in power domain increases, the multiple access interference (MAI) becomes severe and the performance of NOMA degrades.

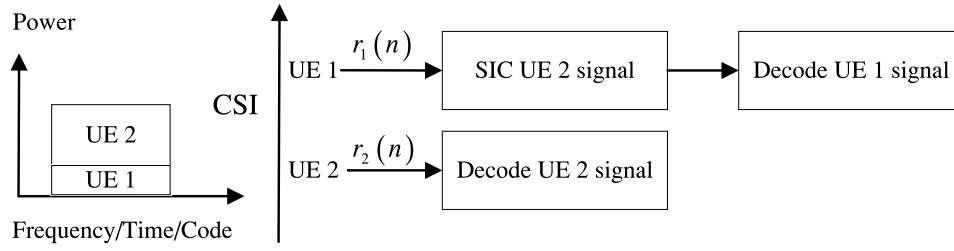


Fig. 8. Downlink power-domain non-orthogonal multiple access (NOMA).

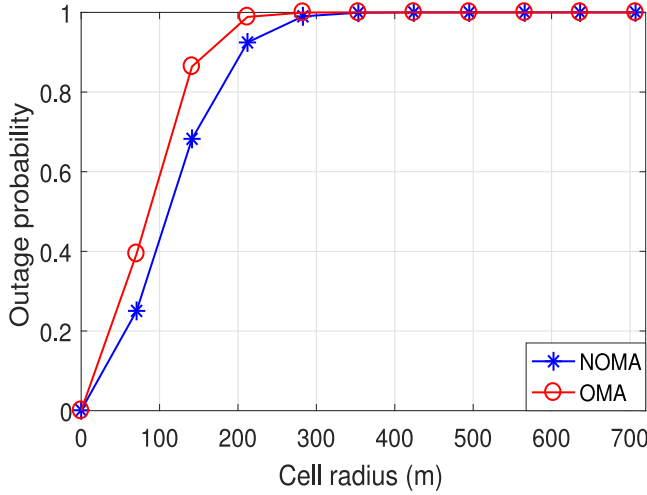


Fig. 9. Outage probability of NOMA and OMA with different cell radius, two users are supported by NOMA scheme where the far user is located at the edge of the cell, and the one user supported by OMA scheme is randomly located in the cell. The path loss exponent is 2. The transmit power signal-to-noise (SNR) is 40 dB. Power allocation coefficients for two users in NOMA are 4/5 and 1/5, respectively.

2) *Power Allocation in NOMA*: NOMA is capable of supporting unequal transmission rates for users experiencing various channel conditions by assigning them different transmit powers. Therefore, the power allocation mechanism for different users is critical to power-domain NOMA. As aforementioned, the SIC receiver works in the descending order of the signal strengths. Specifically, more powers are allocated to the users with poor CSI. By doing so, interference from the users with good CSI is reduced significantly as less powers are allocated to them. Meanwhile, the detection accuracy at users with poor CSI can also be improved. As the power allocation in NOMA is based on the order of CSI, the cases with perfect and imperfect CSI are different and should be investigated separately, as in [47]. When perfect CSI is available, the optimization problem can be formulated to maximize the individual/sum rate while considering the fairness among different users. While with average CSI, the optimization problem can be formulated to minimize the maximum outage probability. A comprehensive summary of the state-of-the-art work on the power allocation in NOMA with different fairness strategies is given in Table I. It is worth noting that the most critical challenge for power allocation in NOMA comes from the non-convex property of the ordered power constraints, which makes the optimization problem untraceable. Therefore,

further research work on the optimal ordering design can be expected.

3) *Multiple Antennas Based NOMA*: Multiple antenna techniques can provide an additional degree of freedom on the spatial domain, and bring further performance improvements to NOMA. Recently, multiple antenna based NOMA has attracted lots of attention [39], [52], [62]–[76]. Different from single-input-single-output (SISO) based NOMA, where the channels are normally represented by scalars, one of the research challenges in multiple antenna based NOMA comes from user ordering; as the channels are generally in form of vectors or matrices. Currently, the possible designs of multiple antenna based NOMA fall into two categories where one or multiple users are served by a single beamforming vector.

By allocating different users with different beams in the same resource block, the quality of service (QoS) of each user can be guaranteed in multiple antenna based NOMA systems forcing the beams to satisfy a predefined order. This type of multiple antenna based NOMA scheme has been first proposed by Sun *et al.* [71] to investigate power optimization to maximize the ergodic capacity. This proposed multiple antenna based NOMA scheme has proved to be able to achieve significant performance improvement compared with conventional OMA schemes.

A cluster of users can share the same beam. The spatial channels of different users within the same cluster are considered to be highly correlated. Therefore, beams for different clusters should be carefully designed to guarantee that the channels for different clusters are orthogonal to each other in order to suppress the inter-cluster interference. For multiple-input-single-output (MISO) based NOMA, a two-stage multicast beamforming scheme has been proposed by Choi [62], where ZF beamforming has been employed to mitigate interference from adjacent clusters first and then the optimal beamforming vectors have been designed to minimize the total transmit power within each cluster. For MIMO based NOMA, a scheme to simultaneously apply open-loop random beamforming and intra-beam SIC, has been proposed by Higuchi and Kishiyama [39]. However, here the system performance is considerably degraded as the random beamforming can bring uncertainties at the user side. More recently, a precoding and detection framework with fixed power allocation has been proposed by Ding *et al.* [65] to solve these problems caused by random beamforming, and demonstrated that MIMO based NOMA can achieve better outage performance than MIMO based OMA even for users who experience strong co-channel interference. Additionally, Zeng *et al.* [72], [73]

TABLE I
POWER ALLOCATION IN NOMA

Power Allocation Strategy	Fairness	Complexity	Ref.
Allocating more power for weak users	Moderate	Low	[48]–[51]
Give absolute priority to the fairness	High	High	[47], [52], [53]
Maximize the geometric mean of user rates	Moderate	Low	[54]–[57]
Give users weighted priorities	Low	Low	[58]
Tradeoff between throughput and fairness	—	—	[59]–[61]

TABLE II
MULTIPLE ANTENNA BASED NOMA

Scenarios	Techniques	Metrics	Characteristics	Ref.
MU-MISO	Two stage BF	Transmit power/rate	Optimal power allocation	[62]
MU-MISO	Restrictive BF	Throughput	Less constraints at the BS	[63]
MU-MISO	ZF-BF + DPC	Power consumption & OP	Precoding with low complexity & user pairing	[64]
MU-MIMO	Random BF	Sum rates	No perfect CSI required	[39]
MU-MIMO	ZF-BF	Max-min rate	Fairness in MIMO-NOMA	[52]
MU-MIMO	ZF-BF	OP+rate gain	No CSI available at the transmitter	[65]
MU-MIMO	Signal alignment	OP	Larger diversity gain	[69]
SU-MIMO	QR based deposition	OP	Channel differences	[70]
SU-MIMO	Beamformer-based	Ergodic capacity	Optimal/sub-optimal power allocation	[71]
MU-MIMO	Two-stage precoding	OP+sum rate	One bit feedback	[74]

have proved the superiority of MIMO-NOMA over MIMO-OMA in terms of both sum channel capacity and ergodic sum capacity. Furthermore, two coordinated beamforming approaches has been proposed by Shin *et al.* [66] to deal with of inter-cell interference in two-cell MIMO-NOMA networks, in which two BSs jointly optimize their beamforming vectors to improve the data rates of cell-edge users. This approach is extended to an arbitrary number of cells in [67], in which the maximum number of users supported by the proposed scheme in multi-cell MIMO networks is characterized. More details about multi-cell MIMO-NOMA approaches can be found in [68].

A comprehensive summary for the state-of-the-art work on multiple antenna based NOMA is given in Table II, where “BF”, “OP”, “SU” and “MU” are used to represent beamforming, outage probability, two-user and multi-users cases, respectively.

4) *Cooperative NOMA*: In cellular networks, a cell-edge user usually experiences a weaker received signal power and lower data rates compared to those near the BS. Relaying and coordinated multipoint (CoMP) transmission (and reception) techniques have been widely employed to increase the transmission rates for cell-edge users [77]. The scenario with users transmitting at different rates naturally matches the application scenarios typical of NOMA.

The basic idea of relay-assisted NOMA is to use the users with the better CSI as the decode-and-forward (DF) or

amplify-and-forward (AF) as relays to improve the transmission rates of the users with poor CSI. A cooperative NOMA model supporting M users with M time slots has been proposed in [78]. In the first time slot, the traditional non-cooperative NOMA scheme is conducted. In the second time slot, the user with the best CSI acts as the DF relay for the user with the second best CSI. In the following time slots, the user with the m -th best CSI works as the relay for the user with the subsequent worse CSI to improve the transmission rates.

CoMP transmission, where multiple BSs support cell-edge users together, is capable of improving the performance of cell-edge users. NOMA has been first applied into CoMP by Choi [79], where two coordinated BSs use Alamouti code to support a cell-edge user in a NOMA channel. Subsequently, the effectiveness of NOMA in CoMP systems has been further demonstrated by Tian *et al.* [80] in comparison with the conventional joint-transmission NOMA. A coordinated direct and relay transmission (CDRT) scheme has also been considered by Kim and Lee [81], where the BS communicates with a near user and a relay simultaneously, invoking NOMA in the first time slot while communicating with a far user with the aid of the relay in the following time slots. This NOMA based CDRT scheme solves the main challenge by using the inherent property of NOMA that allows a receiver to obtain side information such as other user’s signal for interference cancellation.

TABLE III
COOPERATIVE NOMA

Techniques	Metrics	Characteristics	Ref.
AF	OP+ergodic rate	Multiple antenna relay	[82]
DF	OP+throughput	Self-powered relays	[51]
DF	OP+ergodic rate	Maximum diversity gain for all users	[78]
DF	OP	Two stage relay selection schemes	[83]
DF	Sum rate	Novel two-stage power allocation	[84]
CoMP	Sum rate	New coordinated superposition coding	[79]
CoMP+DF	OP+ergodic rate	Each stream has same diversity order	[81]

A comprehensive summary of the existing work on NOMA in cooperative communications is given in Table III, where “OP” represents the outage probability.

5) *Cognitive Radio Inspired NOMA*: To address the spectrum scarcity issue in wireless communications, cognitive ratio (CR) has been proposed [85], [86], which allows unlicensed users to use licensed spectrum as long as they generate no (intolerable) interference to the licensed users. Spectrum sensing [87]–[89] is an enabling technique for CR, and how to balance the sensing overhead and CR throughput has been investigated in [90] and [91].

If an unlicensed user is at a distance away from the licensed user that shares the same licensed spectrum, then the interference between them will be small in general. Therefore, CR in this case can be regarded as special NOMA in the location domain while the NOMA introduced in the previous sections is in power, code, or multiple domains. Device-to-device (D2D) communications [92], [93], a popular research area recently, turn to a special CR if the two users in D2D communications are regarded as a simple CR network. In addition to the location domain, CR can also exploit the spatial domain [94], [95] and the frequency-spatial domain [96].

The network model of underlay CR is similar to NOMA, which allocates power of different users under some SINR constraints. Therefore, a cognitive-radio-inspired NOMA (CR-NOMA) scheme has been proposed in [97], which is different from the traditional scheme with fixed power allocation. In [98], NOMA-based underlay CR networks have been studied and the closed-form expressions of the outage probability have been derived. In addition, in [99]–[102], power allocation schemes for CR-NOMA and NOMA-based CR networks have been developed to improve the system performance in terms of sum rate, energy efficiency (EE), etc.

6) *SE and EE in NOMA*: Recall that when designing 5G networks, SE and EE are two important performance metrics. NOMA provides a promising solution due to its extra degree of freedom over the power domain, especially suited for IoT networks that require massive connectivity but with low power consumption at sensor nodes. More specifically, NOMA has been considered to be able to boost the SE of 5G networks [81], [98], [103], [104]. NOMA with the consideration of EE has also been investigated in [105], where a

sub-optimal resource allocation algorithm has been proposed to maximize the EE of the systems. Moreover, the EE of a NOMA network, subject to a minimum required data rate for each user, can be maximized using the approach in [106]. Similar to most wireless networks [107], SE and EE cannot be achieved simultaneously in NOMA networks. Therefore we expect to see more work on SE and EE tradeoffs in NOMA in the future. A more comprehensive review on the power-domain NOMA can be found in [108].

B. Code-Domain NOMA

Code-domain NOMA can support multiple transmissions within the same time-frequency resource block by assigning different codes to different users. It has certain spreading gain and shaping gain with the cost of extra signal bandwidth in comparison with power-domain NOMA. Existing solutions to code-domain NOMA mainly include low-density spreading CDMA (LDS-CDMA) [109], low-density spreading OFDM (LDS-OFDM) [59], and sparse code multiple access (SCMA) [110], which will be introduced in the following.

1) *LDS-CDMA*: LDS-CDMA [109] is a novel type of CDMA. Its key feature is that a low-density signature, which has a similar form of the low-density parity-check (LDPC) matrix, is employed for the codebook construction. When the number of users is larger than that of samples per symbol period in the conventional CDMA, MAI is inevitable and optimal multiuser detection is extremely complex. However, due to the sparse structure of the signature in LDS-CDMA, a low-complexity near-optimal multiuser detection scheme, based on a message passing algorithm (MPA), can be applied in the detection of LDS-CDMA, which significantly improves performance.

2) *LDS-OFDM*: LDS-OFDM [59] has similar properties to LDS-CDMA, except that the output of the signature is mapped into the subcarriers of OFDM rather than the time samples in CDMA. Therefore, a low-complexity MPA detector can be adopted. Compared to LDS-CDMA, LDS-OFDM utilizes multicarrier transmission, which makes it fit for wide-band channels. Further, the strong compatibility with OFDM makes it flexible in resource allocation [59].

3) *SCMA*: In SCMA [110], by applying a sparse code book similar to the signature matrix in LDS, a certain number

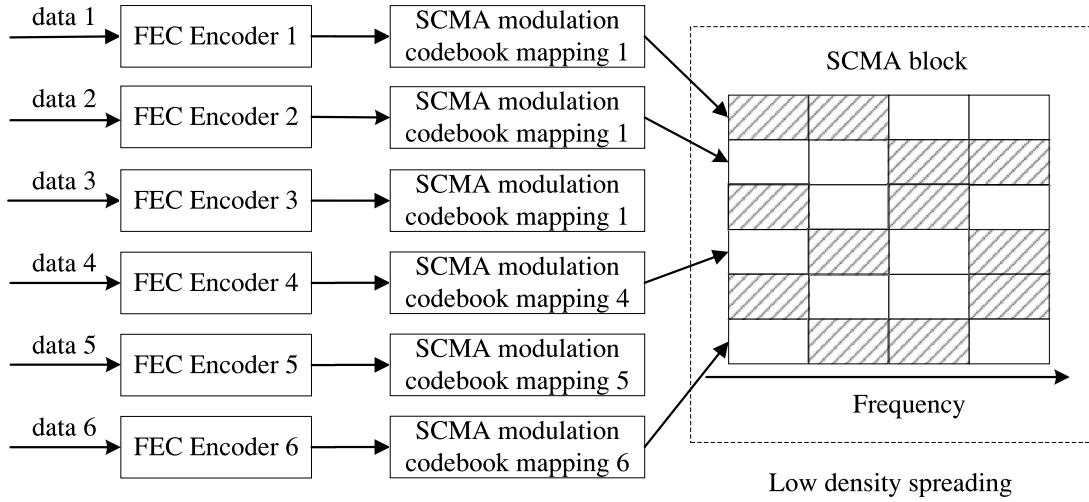


Fig. 10. Sparse code multiple access (SCMA).

of resource blocks can support more users through spreading. Fig. 10 demonstrates SCMA where six users share four resource blocks. For the example in Fig. 10, the signature matrix can be expressed as

$$\mathbf{S} = \begin{bmatrix} 1 & 0 & 1 & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 & 0 & 1 \\ 0 & 1 & 1 & 0 & 0 & 1 \\ 0 & 1 & 0 & 1 & 1 & 0 \end{bmatrix}. \quad (13)$$

Although a part of the users share the same block, another block would be adopted to distinguish different users when collisions occur.

Besides the sparse spreading, SCMA utilizes multi-dimensional constellations to reduce the receiver complexity and further improve the SE. Attributed to the multi-dimension property, the constellation in one resource block can be projected into its subspace [111]. For example, a four-point QAM constellation can be projected to a three-point constellation. Even when two points collide in one resource block or to say one dimension, they can be distinguished in the other used blocks. Due to fewer constellation points, the receiver complexity can be reduced. Moreover, the constellation design can focus on improving the detection performance. For example, a design based on constellation rotation and interleaving has been proposed in [112], which is able to achieve better BER performance compared to the simple LDS-OFDM.

Due to the sparse structure of the spreading matrix and the large minimum distance of the multi-dimensional constellation, the detection performance of SCMA becomes excellent even when the resource blocks are overloaded. At the receiver, MPA, which is usually adopted in the decoding of LDPC, is applied in the detection [113], [114]. Due to the sparsity, MPA could achieve near-optimal performance with a much lower complexity compared to the optimal maximum likelihood (ML) and the BCJR algorithms. However, the complexity is still relatively high for user devices. Hence, SCMA also considers clustering the users based on the CSI and allocating different powers to different clusters. When the transmit powers among different clusters vary, the signals of different

clusters can be detected by using SIC, which is similar to the power-domain NOMA. Within each cluster, different users can be distinguished by using MPA. As a result, the combination of SIC and MPA can reduce the complexity of the receiver significantly.

C. NOMA Multiplexing in Multiple Domains

Beyond multiplexing signals in power domain or code domain, some of solutions for NOMA have been proposed to multiplex in multiple domains, such as the power domain, the code domain, and the spatial domain, in order to support massive connectivity for 5G networks. In Section III-A2, we discussed multiple antenna based NOMA, where NOMA multiplexed in the power and spatial domains. We now introduce another three types of typical NOMA schemes multiplexing in multiple domains: pattern division multiple access (PDMA) [115], building block sparse-constellation based orthogonal multiple access (BOMA) [116], and lattice partition multiple access (LPMA) [117].

1) PDMA: In PDMA [115], non-orthogonal patterns are allocated to different users to perform multiplexing. These patterns are carefully designed in the multiple domains of code, power, and space, to gain the SIC-amenable property. In the presence of this property, the low-complexity SIC based MPA multiuser detection method with reliable performance can be designed to run at the receiver side.

At the transmitter, similar to SCMA, the users in PDMA are also spread by a sparse signature matrix [115]. The main difference is that the number of resource blocks occupied by each user in PDMA can vary. For example, seven users can be multiplexed within three resource blocks through the following signature matrix

$$\mathbf{S} = \begin{bmatrix} 1 & \sqrt{\frac{3}{2}} & \sqrt{\frac{3}{2}} & 0 & \sqrt{3} & 0 & 0 \\ 1 & \sqrt{\frac{3}{2}} & 0 & \sqrt{\frac{3}{2}} & 0 & \sqrt{3} & 0 \\ 1 & 0 & \sqrt{\frac{3}{2}} & \sqrt{\frac{3}{2}} & 0 & 0 & \sqrt{3} \end{bmatrix}. \quad (14)$$

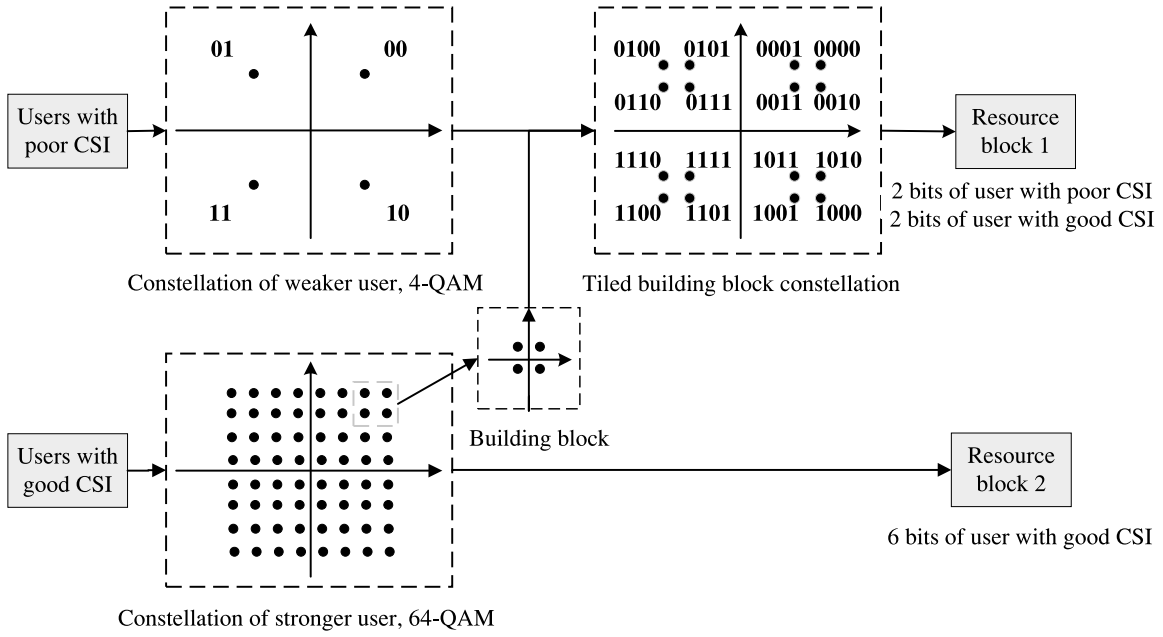


Fig. 11. Building block sparse-constellation based orthogonal multiple access (BOMA).

By utilizing the sparse signature matrix, PDMA can increase the system capacity through overloading. Moreover, users can also be multiplexed in other domains, such as power and space. In the same resource block, users can be distinguished by different powers as the power-domain NOMA or different precoders if MIMO is applied [118].

At the receiver side, similar to SCMA, MPA can be adopted in detection due to the sparsity of the signature matrix. When different clusters of users are multiplexed in power and space domains, MPA-SIC can be applied. The detection of users that are multiplexed in the same signature matrix is based on the MPA, which can provide excellent performance. Among different clusters in the power and space domains, SIC can be utilized to reduce the complexity. Besides, a turbo structure can be adopted to combine the detector with the decoder to further improve the performance [119].

2) *BOMA*: This technique attaches the information from a user with good CSI to the symbols of a user with poor CSI. Thus the capacity of a multiuser system is increased significantly.

As shown in Fig. 11, in order to achieve the same BER performance as a user with good CSI, the user with poor CSI should apply a coarse constellation with a large minimum distance. Hence, the small building block that contains the data of the user with good CSI can be tiled in the constellation of the user with poor CSI [120], [121]. For the user with poor CSI, the center of the building block can be regarded as the constellation point and the tiled building block can be regarded as interference. When the size of the building block is much smaller than the minimum distance of the coarse constellation, the degradation of detection performance becomes minimal. Since the user with good CSI can detect the points in its own constellation, it can also detect all points in the tiled building block constellation and decode the bits from itself.

The structure of BOMA is simple and similar to that adopted in current 4G systems. Only minor software changes are required so that BOMA can be easily implemented with the compatibility to massive MIMO, high frequency bands, and other requirements of 5G systems. Besides, BOMA needs no complex power allocation and SIC receiver that are necessary for other NOMA schemes.

3) *LPMA*: In LPMA [117], the power domain and code domain are combined to multiplex users. Similar to power multiplexing in power-domain NOMA, the code in LPMA implements a multilevel lattice that allocates different code levels to users with different CSI. Several types of codes can be adopted, such as Construction π_A [122] and Construction D [123]. For users with the poor CSI, the allocated codes have larger minimum distance that can improve detection performance. For users with better CSI, the allocated codes are with smaller minimum distance without degrading detection performance. At the receiver, a SIC decoder is adopted, which is similar to that found in power-domain NOMA.

Besides the code domain multiplexing, LPMA also adopts power multiplexing to enhance those users with poor CSI. With the aid of two degrees of freedom in the multiplexing, the design of LPMA becomes more flexible in comparison with power-domain NOMA. Even if a pair of users have similar CSI, they can still be multiplexed by adjusting the allocated code levels and power levels, therefore the complex user clustering mechanisms adopted in the power-domain NOMA schemes are not required in LPMA.

D. Comparison

In the above, we have introduced several typical NOMA schemes based on various multiplexing techniques. With different multiplexing techniques, the features of these NOMA schemes are different. The power-domain NOMA has a simple

TABLE IV
COMPARISONS OF NOMA SCHEMES

Schemes	Characteristics	Advantages	Disadvantages
Power-domain NOMA	Power domain multiplexing	High SE Compatible to other techniques	Need user pairing Error propagation in SIC
LDS-CDMA	Sparse spreading CDMA	No need of CSI Near-optimal MPA detector	Redundancy from coding
LDS-OFDM	Sparse spreading OFDM	No need of CSI Near-optimal MPA detector More fit for wideband than LDS-CDMA	Redundancy from coding
SCMA	Sparse spreading Multi-dimensional constellation	No need of CSI Near-optimal MPA detector More diversity than simple LDS	Redundancy from coding Difficult to design optimal codebook
PDMA	Sparse spreading Multiplexing in power, code, and spatial domains	More diversity Near-optimal MPA detector Low-complexity receiver	Redundancy from coding Difficult to design optimal patterns
BOMA	Tiled building block	Simple structure Compatible to current system Low-complexity receiver	Need user pairing Not very flexible
LPMA	Multilevel lattice code Multiplexing in power and code domains	No need for user clustering	Specific channel coding

structure and can be easily combined with various technologies, such as MIMO and cooperative networks. However, user clustering and pairing are required to get the user order in terms of CSI, which increases the system complexity. Code-domain NOMA and the NOMA multiplexing in multiple domains, such as SCMA and PDMA, can obtain the spreading gain without requiring perfect CSI. Besides, the low-complexity near-optimal MPA detectors can achieve better performance than SIC detectors. However, the coding also introduces redundancy, which degrades the system SE. Other NOMA schemes also have their specific features. For example, BOMA has a simple structure and can be easily compatible to the current LTE systems while user pairing is also required, which reduces the flexibility. LPMA can achieve multilevel in both the power domain and the code domain, which saves the cost of user clustering. However, the multilevel code requires corresponding non-binary or nested binary channel coding. The key techniques and features of the aforementioned NOMA schemes are summarized in Table IV. It is noted that each NOMA scheme has its advantages and limitations, which fits different application situations. Actually, it is an adaptive configuration to realize the trade-off between performance and implementation complexity. For example, if there exists a large difference among users' channel conditions due to the near-far effect or in moving networks, power-domain NOMA with a SIC receiver can be used with relatively low complexity. On the other hand, if the application scenarios require high

reliability, especially when channel condition is bad or the location distribution of users is concentrated, SCMA is a feasible solution due to its shaping gain and near-optimal MPA detection.

E. Future Work

Several NOMA schemes have been discussed in this section. Even if using different techniques, these schemes share the same spirit to utilize non-orthogonality to increase the system capacity and support more users by the limited resource blocks. Beyond the existing work, more research is necessary to improve the performance of these NOMA schemes from the following aspects.

- 1) The MPA-SIC detection method is usually applied in SCMA and PDMA, in which the user clustering mechanism affects the performance of the method significantly. When users are asynchronous, those with similar time delays should be divided into the same cluster for better performance. If the delays vary a lot among the users within the same cluster, interference among different users becomes large and may break the sparse structure. Multi-branch technique [124] can be applied to improve the performance by regarding each cluster as a branch. By calculating each branch in parallel and selecting the best result as the final one, the performance could be improved compared to the single clustering approach.

- 2) The joint design of new modulation and NOMA schemes is an important direction to be explored in 5G networks. Some of the NOMA schemes, especially the LDS based code-domain NOMA, are based on OFDM, where the output of the sparse spreading matrix is mapped into orthogonal subcarriers. In general, how to properly combine the modulation and NOMA scheme is under research. For example, for the combination of SCMA and f-OFDM, the short CP of f-OFDM could introduce ISI and ICI when the subband is narrow and degrade the detection performance of SCMA. If the RISIC algorithm is adopted to cancel the interference introduced by f-OFDM, the multiuser detection of SCMA should be included in the iteration of CP reconstruction, which poses a requirement of joint design approaches for the receivers.
- 3) Despite the extensive efforts on NOMA, most of existing studies focus on the performance analysis and algorithm design of various NOMA techniques individually, such as power-domain NOMA and code-domain NOMA. However, different scenarios have their preferred NOMA techniques, for instance, if users experience very different channel conditions due to the near-far effect, power-domain NOMA can be a better candidate. If users experience poor channel conditions but require high reliability, SCMA is preferred due to its shaping gain and near-optimal MPA detection. Qin *et al.* [125] has proposed a unified NOMA framework including both power and code domains to support various scenarios by the same hardware infrastructure in heterogeneous ultra dense networks. However, more research efforts are still desired to realize such a unified NOMA framework.
- 4) The design of modulation and MA schemes for high frequency bands (above 40 GHz) is beginning to receive increased interest. The millimeter-wave (mmWave) and Terahertz (THz) bands appear to be good candidates to decrease spectrum scarcity due to the availabilities in current circuit design [126], [127]. However, the propagation properties of mmWave and THz bands have shown to be quite poor, which brings new challenges on system designs. For example, noise is the major limitation of mmWave and THz bands, which makes the transmit power levels extremely important and ultimately impacts the classes of applications that can use them, e.g., IoT. Moreover, high level impairments including carrier frequency offset (CFO) and phase noise also need to be considered in mmWave and THz bands as they are noise-limited. Nevertheless, there is already a study on NOMA based mmWave communications [128], we may further see analyses of such systems based on practical scenarios in the future.

IV. CONCLUSION

In this article, we provide a comprehensive survey covering the major promising candidates for modulation and multiple access (MA) in fifth generation (5G) networks. From our discussion, we can see that new modulations for orthogonal MA

can be adopted to reduce out-of-band leakage while meeting the diverse demands of 5G networks. Non-orthogonal MA is another promising approach that marks a deviation from the previous generations of wireless networks. By utilizing non-orthogonality, we have convincingly shown that 5G networks will be able to provide enhanced throughput and massive connectivity with improved spectral efficiency.

REFERENCES

- [1] "Study on scenarios and requirements for next generation access technologies," 3GPP, Sophia Antipolis, France, Tech. Rep. 38.913, Jun. 2017.
- [2] "IMT vision—Framework and overall objectives of the future development of IMT for 2020 and beyond," Int. Telecommun. Union, Geneva, Switzerland, Recommendation ITU-R M.2083, Sep. 2015.
- [3] H. Sampath, S. Talwar, J. Tellado, V. Erceg, and A. Paulraj, "A fourth-generation MIMO-OFDM broadband wireless system: Design, performance, and field trial results," *IEEE Commun. Mag.*, vol. 40, no. 9, pp. 143–149, Sep. 2002.
- [4] Y. G. Li and G. Stuber, *OFDM for Wireless Communications*. Boston, MA, USA: Springer, 2006.
- [5] G. L. Stuber *et al.*, "Broadband MIMO-OFDM wireless communications," *Proc. IEEE*, vol. 92, no. 2, pp. 271–294, Feb. 2004.
- [6] V. Vakilian, T. Wild, F. Schaich, S. T. Brink, and J.-F. Frigon, "Universal-filtered multi-carrier technique for wireless systems beyond LTE," in *Proc. IEEE GLOBECOM Workshops (GC Wkshps)*, Atlanta, GA, USA, Dec. 2013, pp. 223–228.
- [7] F. Schaich, T. Wild, and Y. Chen, "Waveform contenders for 5G—Suitability for short packet and low latency transmissions," in *Proc. IEEE Veh. Technol. Conf. (VTC Spring)*, Seoul, South Korea, May 2014, pp. 1–5.
- [8] J. Abdoli, M. Jia, and J. Ma, "Filtered OFDM: A new waveform for future wireless systems," in *Proc. IEEE Int. Workshop Signal Process. Adv. Wireless Commun. (SPAWC)*, Stockholm, Sweden, Jun. 2015, pp. 66–70.
- [9] X. Zhang, M. Jia, L. Chen, J. Ma, and J. Qiu, "Filtered-OFDM—Enabler for flexible waveform in the 5th generation cellular networks," in *Proc. IEEE Glob. Commun. Conf. (GLOBECOM)*, San Diego, CA, USA, Dec. 2015, pp. 1–6.
- [10] B. Farhang-Boroujeny, "OFDM versus filter bank multicarrier," *IEEE Signal Process. Mag.*, vol. 28, no. 3, pp. 92–112, May 2011.
- [11] M. Bellanger, M. Renfors, T. Ihalainen, and C. A. F. da Rocha, "OFDM and FBMC transmission techniques: A compatible high performance proposal for broadband power line communications," in *Proc. Int. Symp. Power Line Commun. Appl. (ISPLC)*, Rio de Janeiro, Brazil, Mar. 2010, pp. 154–159.
- [12] G. Fettweis, M. Krondorf, and S. Bittner, "GFDM—Generalized frequency division multiplexing," in *Proc. IEEE Veh. Technol. Conf. (VTC Spring)*, Barcelona, Spain, Apr. 2009, pp. 1–4.
- [13] N. Michailow *et al.*, "Generalized frequency division multiplexing for 5th generation cellular networks," *IEEE Trans. Commun.*, vol. 62, no. 9, pp. 3045–3061, Sep. 2014.
- [14] B. Boashash, Ed., *Time Frequency Signal Analysis and Processing: A Comprehensive Reference*. Oxford, U.K.: Elsevier, 2003.
- [15] C.-D. Chung, "Spectrally precoded OFDM," *IEEE Trans. Commun.*, vol. 54, no. 12, pp. 2173–2185, Dec. 2006.
- [16] J. van de Beek and F. Berggren, "N-continuous OFDM," *IEEE Commun. Lett.*, vol. 13, no. 1, pp. 1–3, Jan. 2009.
- [17] Y. Zheng, J. Zhong, M. Zhao, and Y. Cai, "A precoding scheme for N-continuous OFDM," *IEEE Commun. Lett.*, vol. 16, no. 12, pp. 1937–1940, Dec. 2012.
- [18] J. van de Beek, "Sculpting the multicarrier spectrum: A novel projection precoder," *IEEE Commun. Lett.*, vol. 13, no. 12, pp. 881–883, Dec. 2009.
- [19] A. Monk, R. Hadani, M. Tsatsanis, and S. Rakib, "OTFS—Orthogonal time frequency space," *arXiv preprint*, Aug. 2016. [Online]. Available: <http://arxiv.org/abs/1608.02993>
- [20] M. Agiwal, A. Roy, and N. Saxena, "Next generation 5G wireless networks: A comprehensive survey," *IEEE Commun. Surveys Tuts.*, vol. 18, no. 3, pp. 1617–1655, 3rd Quart., 2016.
- [21] X. Zhang, L. Chen, J. Qiu, and J. Abdoli, "On the waveform for 5G," *IEEE Commun. Mag.*, vol. 54, no. 11, pp. 74–80, Nov. 2016.

- [22] B. Farhang-Boroujeny and H. Moradi, "OFDM inspired waveforms for 5G," *IEEE Commun. Surveys Tuts.*, vol. 18, no. 4, pp. 2474–2492, 4th Quart., 2016.
- [23] Y. Wang, B. Ren, S. Sun, S. Kang, and X. Yue, "Analysis of non-orthogonal multiple access for 5G," *China Commun.*, vol. 13, no. 2, pp. 52–66, 2016.
- [24] L. Dai *et al.*, "Non-orthogonal multiple access for 5G: Solutions, challenges, opportunities, and future research trends," *IEEE Commun. Mag.*, vol. 53, no. 9, pp. 74–81, Sep. 2015.
- [25] Z. Ding *et al.*, "Application of non-orthogonal multiple access in LTE and 5G networks," *IEEE Commun. Mag.*, vol. 55, no. 2, pp. 185–191, Feb. 2017.
- [26] T. Yunzheng, L. Long, L. Shang, and Z. Zhi, "A survey: Several technologies of non-orthogonal transmission for 5G," *China Commun.*, vol. 12, no. 10, pp. 1–15, Oct. 2015.
- [27] S. Zhang, X. Xu, Y. Wu, L. Lu, and Y. Chen, "A survey on 5G new waveform: From energy efficiency aspects," in *Proc. Asilomar Conf. Signals Syst. Comput.*, Pacific Grove, CA, USA, Nov. 2014, pp. 1939–1943.
- [28] A. Sahin, I. Guvenc, and H. Arslan, "A survey on multicarrier communications: Prototype filters, lattice structures, and implementation aspects," *IEEE Commun. Surveys Tuts.*, vol. 16, no. 3, pp. 1312–1338, 3rd Quart., 2014.
- [29] "Transmit/receive processing (single antenna), deliverable 2.1," document ICT-211887 PHYDYAS, Jul. 2008.
- [30] Z. Zhao *et al.*, "Pulse shaped OFDM for asynchronous uplink access," in *Proc. Asilomar Conf. Signals Syst. Comput.*, Monterey, CA, USA, Nov. 2015, pp. 3–7.
- [31] C. Kim *et al.*, "QAM-FBMC: A new multi-carrier system for post-OFDM wireless communications," in *Proc. IEEE Glob. Commun. Conf. (GLOBECOM)*, San Diego, CA, USA, Dec. 2015, pp. 1–6.
- [32] J. Li, K. Kearney, E. Bala, and R. Yang, "A resource block based filtered OFDM scheme and performance comparison," in *Proc. Int. Conf. Telecommun. (ICT)*, Casablanca, Morocco, May 2013, pp. 1–5.
- [33] U. Kumar, C. Ibars, A. Bhorkar, and H. Jung, "A waveform for 5G: Guard interval DFT-s-OFDM," in *Proc. IEEE GLOBECOM Workshops (GC Wkshps)*, San Diego, CA, USA, Dec. 2015, pp. 1–6.
- [34] G. Berardinelli, F. M. L. Tavares, T. B. Sørensen, P. Mogensen, and K. Pajukoski, "Zero-tail DFT-spread-OFDM signals," in *Proc. IEEE GLOBECOM Workshops (GC Wkshps)*, Atlanta, GA, USA, Dec. 2013, pp. 229–234.
- [35] A. Sahin, R. Yang, M. Ghosh, and R. L. Olesen, "An improved unique word DFT-spread OFDM scheme for 5G systems," in *Proc. IEEE GLOBECOM Workshops (GC Wkshps)*, San Diego, CA, USA, Dec. 2015, pp. 1–6.
- [36] P. Achaichia, M. L. Bot, and P. Siohan, "Windowed OFDM versus OFDM/OQAM: A transmission capacity comparison in the HomePlug AV context," in *Proc. Int. Symp. Power Line Commun. Appl. (ISPLC)*, Udine, Italy, pp. 405–410, Apr. 2011.
- [37] D. Kim and G. L. Stuber, "Residual ISI cancellation for OFDM with applications to HDTV broadcasting," *IEEE J. Sel. Areas Commun.*, vol. 16, no. 8, pp. 1590–1599, Oct. 1998.
- [38] X. Li and L. J. Cimini, "Effects of clipping and filtering on the performance of OFDM," *IEEE Commun. Lett.*, vol. 2, no. 5, pp. 131–133, May 1998.
- [39] K. Higuchi and Y. Kishiyama, "Non-orthogonal access with random beamforming and intra-beam SIC for cellular MIMO downlink," in *Proc. IEEE Veh. Technol. Conf. (VTC Fall)*, Las Vegas, NV, USA, Sep. 2013, pp. 1–5.
- [40] "TP for classification of MUST schemes," 3GPP, Sophia Antipolis, France, Tech. Rep. R1-154999, Aug. 2015.
- [41] *Downlink Multiuser Superposition Transmissions for LTE*, document RP-160680, 3GPP, Sophia Antipolis, France, Mar. 2016.
- [42] A. Duel-Hallen, J. Holtzman, and Z. Zvonar, "Multiuser detection for CDMA systems," *IEEE Pers. Commun.*, vol. 2, no. 2, pp. 46–58, Apr. 1995.
- [43] S. Moshavi, "Multi-user detection for DS-CDMA communications," *IEEE Commun. Mag.*, vol. 34, no. 10, pp. 124–136, Oct. 1996.
- [44] S. Verdú, "Optimum multiuser asymptotic efficiency," *IEEE Trans. Commun.*, vol. 34, no. 9, pp. 890–897, Sep. 1986.
- [45] W.-H. Sheen and G. L. Stuber, "MLSE equalization and decoding for multipath-fading channels," *IEEE Trans. Commun.*, vol. 39, no. 10, pp. 1455–1464, Oct. 1991.
- [46] A. J. Viterbi, "Very low rate convolution codes for maximum theoretical performance of spread-spectrum multiple-access channels," *IEEE J. Sel. Areas Commun.*, vol. 8, no. 4, pp. 641–649, May 1990.
- [47] S. Timotheou and I. Krikidis, "Fairness for non-orthogonal multiple access in 5G systems," *IEEE Signal Process. Lett.*, vol. 22, no. 10, pp. 1647–1651, Oct. 2015.
- [48] Y. Saito *et al.*, "Non-orthogonal multiple access (NOMA) for cellular future radio access," in *Proc. IEEE Veh. Technol. Conf. (VTC Spring)*, Dresden, Germany, Jun. 2013, pp. 1–5.
- [49] Y. Saito, A. Benjebbour, Y. Kishiyama, and T. Nakamura, "System-level performance evaluation of downlink non-orthogonal multiple access (NOMA)," in *Proc. IEEE Int. Symp. Pers. Indoor Mobile Radio Commun. (PIMRC)*, London, U.K., Sep. 2013, pp. 611–615.
- [50] Z. Ding, Z. Yang, P. Fan, and H. V. Poor, "On the performance of non-orthogonal multiple access in 5G systems with randomly deployed users," *IEEE Signal Process. Lett.*, vol. 21, no. 12, pp. 1501–1505, Dec. 2014.
- [51] Y. Liu, Z. Ding, M. Elkashlan, and H. V. Poor, "Cooperative non-orthogonal multiple access with simultaneous wireless information and power transfer," *IEEE J. Sel. Areas Commun.*, vol. 34, no. 4, pp. 938–953, Apr. 2016.
- [52] Y. Liu, M. Elkashlan, Z. Ding, and G. K. Karagiannidis, "Fairness of user clustering in MIMO non-orthogonal multiple access systems," *IEEE Commun. Lett.*, vol. 20, no. 7, pp. 1465–1468, Jul. 2016.
- [53] J. Cui, Z. Ding, and P. Fan, "A novel power allocation scheme under outage constraints in NOMA systems," *IEEE Signal Process. Lett.*, vol. 23, no. 9, pp. 1226–1230, Sep. 2016.
- [54] F. Liu, P. Mähönen, and M. Petrova, "Proportional fairness-based user pairing and power allocation for non-orthogonal multiple access," in *Proc. IEEE Int. Symp. Pers. Indoor Mobile Radio Commun. (PIMRC)*, Hong Kong, Aug. 2015, pp. 1127–1131.
- [55] F. Liu, P. Mähönen, and M. Petrova, "Proportional fairness-based power allocation and user set selection for downlink NOMA systems," in *Proc. Conf. Commun. (ICC)*, Kuala Lumpur, Malaysia, May 2016, pp. 1–6.
- [56] J. Mei, L. Yao, H. Long, and K. Zheng, "Joint user pairing and power allocation for downlink non-orthogonal multiple access systems," in *Proc. Conf. Commun. (ICC)*, Kuala Lumpur, Malaysia, May 2016, pp. 1–6.
- [57] N. Otao, Y. Kishiyama, and K. Higuchi, "Performance of non-orthogonal access with SIC in cellular downlink using proportional fair-based resource allocation," in *Proc. Int. Symp. Wireless Commun. Syst. (ISWCS)*, Paris, France, Aug. 2012, pp. 476–480.
- [58] Y. Sun, D. W. K. Ng, Z. Ding, and R. Schober, "Optimal joint power and subcarrier allocation for MC-NOMA systems," in *Proc. IEEE Glob. Commun. Conf. (GLOBECOM)*, Washington, DC, USA, 2016, pp. 1–6.
- [59] M. Al-Imari, P. Xiao, M. A. Imran, and R. Tafazolli, "Uplink non-orthogonal multiple access for 5G wireless networks," in *Proc. Int. Symp. Wireless Commun. Syst. (ISWCS)*, Barcelona, Spain, Aug. 2014, pp. 781–785.
- [60] L. Lei, D. Yuan, C. K. Ho, and S. Sun, "Power and channel allocation for non-orthogonal multiple access in 5G systems: Tractability and computation," *IEEE Trans. Wireless Commun.*, vol. 15, no. 12, pp. 8580–8594, Dec. 2016.
- [61] P. D. Diamantoulakis, K. N. Pappi, Z. Ding, and G. K. Karagiannidis, "Wireless-powered communications with non-orthogonal multiple access," *IEEE Trans. Wireless Commun.*, vol. 15, no. 12, pp. 8422–8436, Dec. 2016.
- [62] J. Choi, "Minimum power multicast beamforming with superposition coding for multiresolution broadcast and application to NOMA systems," *IEEE Trans. Commun.*, vol. 63, no. 3, pp. 791–800, Mar. 2015.
- [63] K. Higuchi and A. Benjebbour, "Non-orthogonal multiple access (NOMA) with successive interference cancellation for future radio access," *IEICE Trans. Commun.*, vol. 98, no. 3, pp. 403–414, Mar. 2015.
- [64] Z. Chen, Z. Ding, X. Dai, and G. K. Karagiannidis, "On the application of quasi-degradation to MISO-NOMA downlink," *IEEE Trans. Signal Process.*, vol. 64, no. 23, pp. 6174–6189, Dec. 2016.
- [65] Z. Ding, F. Adachi, and H. V. Poor, "The application of MIMO to non-orthogonal multiple access," *IEEE Trans. Wireless Commun.*, vol. 15, no. 1, pp. 537–552, Jan. 2016.
- [66] W. Shin *et al.*, "Coordinated beamforming for multi-cell MIMO-NOMA," *IEEE Commun. Lett.*, vol. 21, no. 1, pp. 84–87, Jan. 2017.
- [67] W. Shin, M. Vaezi, J. Lee, and H. V. Poor, "On the number of users served in MIMO-NOMA cellular networks," in *Proc. Int. Symp. Wireless Commun. Syst. (ISWCS)*, Poznan, Poland, Sep. 2016, pp. 638–642.

- [68] W. Shin *et al.*, "Non-orthogonal multiple access in multi-cell networks: Theory, performance, and practical challenges," *IEEE Commun. Mag.*, vol. 55, no. 10, pp. 176–183, Oct. 2017.
- [69] Z. Ding, R. Schober, and H. V. Poor, "A general MIMO framework for NOMA downlink and uplink transmission based on signal alignment," *IEEE Trans. Wireless Commun.*, vol. 15, no. 6, pp. 4438–4454, Jun. 2016.
- [70] Z. Ding, L. Dai, and H. V. Poor, "MIMO-NOMA design for small packet transmission in the Internet of Things," *IEEE Access*, vol. 4, pp. 1393–1405, 2016.
- [71] Q. Sun, S. Han, C.-L. I, and Z. Pan, "On the ergodic capacity of MIMO NOMA systems," *IEEE Wireless Commun. Lett.*, vol. 4, no. 4, pp. 405–408, Aug. 2015.
- [72] M. Zeng, A. Yadav, O. A. Dobre, G. I. Tsiropoulos, and H. V. Poor, "On the sum rate of MIMO-NOMA and MIMO-OMA systems," *IEEE Wireless Commun. Lett.*, vol. 6, no. 4, pp. 534–537, Aug. 2017.
- [73] M. Zeng, A. Yadav, O. A. Dobre, G. I. Tsiropoulos, and H. V. Poor, "Capacity comparison between MIMO-NOMA and MIMO-OMA with multiple users in a cluster," *IEEE J. Sel. Areas Commun.*, vol. 35, no. 10, pp. 2413–2424, Oct. 2017.
- [74] Z. Ding and H. V. Poor, "Design of massive-MIMO-NOMA with limited feedback," *IEEE Signal Process. Lett.*, vol. 23, no. 5, pp. 629–633, May 2016.
- [75] Z. Qin, Y. Liu, Z. Ding, Y. Gao, and M. ElKashlan, "Physical layer security for 5G non-orthogonal multiple access in large-scale networks," in *Proc. Conf. Commun. (ICC)*, Kuala Lumpur, Malaysia, May 2016, pp. 1–6.
- [76] Y. Liu, Z. Qin, M. ElKashlan, Y. Gao, and L. Hanzo, "Enhancing the physical layer security of non-orthogonal multiple access in large-scale networks," *IEEE Trans. Wireless Commun.*, vol. 16, no. 3, pp. 1656–1672, Mar. 2017.
- [77] D. Lee *et al.*, "Coordinated multipoint transmission and reception in LTE-advanced: Deployment scenarios and operational challenges," *IEEE Commun. Mag.*, vol. 50, no. 2, pp. 148–155, Feb. 2012.
- [78] Z. Ding, M. Peng, and H. V. Poor, "Cooperative non-orthogonal multiple access in 5G systems," *IEEE Commun. Lett.*, vol. 19, no. 8, pp. 1462–1465, Aug. 2015.
- [79] J. Choi, "Non-orthogonal multiple access in downlink coordinated two-point systems," *IEEE Commun. Lett.*, vol. 18, no. 2, pp. 313–316, Feb. 2014.
- [80] Y. Tian, A. R. Nix, and M. Beach, "On the performance of opportunistic NOMA in downlink CoMP networks," *IEEE Commun. Lett.*, vol. 20, no. 5, pp. 998–1001, May 2016.
- [81] J.-B. Kim and I.-H. Lee, "Non-orthogonal multiple access in coordinated direct and relay transmission," *IEEE Commun. Lett.*, vol. 19, no. 11, pp. 2037–2040, Nov. 2015.
- [82] J. Men and J. Ge, "Non-orthogonal multiple access for multiple-antenna relaying networks," *IEEE Commun. Lett.*, vol. 19, no. 10, pp. 1686–1689, Oct. 2015.
- [83] Z. Ding, H. Dai, and H. V. Poor, "Relay selection for cooperative NOMA," *IEEE Wireless Commun. Lett.*, vol. 5, no. 4, pp. 416–419, Aug. 2016.
- [84] W. Duan, M. Wen, Y. Yan, Z. Xiong, and M. H. Lee, "Use of non-orthogonal multiple access in dual-hop relaying," *arXiv preprint*, Apr. 2016. [Online]. Available: <http://arxiv.org/abs/1604.01151>
- [85] J. Mitola and G. Q. Maguire, "Cognitive radio: Making software radios more personal," *IEEE Pers. Commun.*, vol. 6, no. 4, pp. 13–18, Aug. 1999.
- [86] Y.-C. Liang, K.-C. Chen, G. Y. Li, and P. Mahonen, "Cognitive radio networking and communications: An overview," *IEEE Trans. Veh. Technol.*, vol. 60, no. 7, pp. 3386–3407, Sep. 2011.
- [87] J. Ma, G. Y. Li, and B.-H. Juang, "Signal processing in cognitive radio," *Proc. IEEE*, vol. 97, no. 5, pp. 805–823, May 2009.
- [88] Z. Qin, Y. Gao, and C. G. Parini, "Data-assisted low complexity compressive spectrum sensing on real-time signals under sub-Nyquist rate," *IEEE Trans. Wireless Commun.*, vol. 15, no. 2, pp. 1174–1185, Feb. 2016.
- [89] Z. Qin, Y. Gao, M. D. Plumbley, and C. G. Parini, "Wideband spectrum sensing on real-time signals at sub-Nyquist sampling rates in single and cooperative multiple nodes," *IEEE Trans. Signal Process.*, vol. 64, no. 12, pp. 3106–3117, Jun. 2016.
- [90] Y.-C. Liang, Y.-H. Zeng, E. C. Y. Peh, and A. T. Hoang, "Sensing-throughput tradeoff for cognitive radio networks," *IEEE Trans. Wireless Commun.*, vol. 7, no. 4, pp. 1326–1337, Apr. 2008.
- [91] Z. Qin, Y. Liu, Y. Gao, M. ElKashlan, and A. Nallanathan, "Wireless powered cognitive radio networks with compressive sensing and matrix completion," *IEEE Trans. Commun.*, vol. 65, no. 4, pp. 1464–1476, Apr. 2017.
- [92] D.-Q. Feng *et al.*, "Device-to-device communications in underlying cellular networks," *IEEE Trans. Commun.*, vol. 61, no. 8, pp. 3541–3551, Aug. 2013.
- [93] D. Feng *et al.*, "Device-to-device communications in cellular networks," *IEEE Commun. Mag.*, vol. 52, no. 4, pp. 49–55, Apr. 2014.
- [94] S. M. Perlaza, N. Fawaz, S. Lasaulce, and M. Debbah, "From spectrum pooling to space pooling: Opportunistic interference alignment in MIMO cognitive networks," *IEEE Trans. Signal Process.*, vol. 58, no. 7, pp. 3728–3741, Jul. 2010.
- [95] R. Sun and Z.-Q. Luo, "Interference alignment using finite and dependent channel extensions: The single beam case," *IEEE Trans. Inf. Theory*, vol. 61, no. 1, pp. 239–255, Jan. 2015.
- [96] L. Lu, G. Y. Li, A. Maaref, and R. Yao, "Opportunistic transmission exploiting frequency- and spatial-domain degrees of freedom," *IEEE Wireless Commun.*, vol. 21, no. 2, pp. 91–97, Apr. 2014.
- [97] Z. Ding, P. Fan, and H. V. Poor, "Impact of user pairing on 5G nonorthogonal multiple-access downlink transmissions," *IEEE Trans. Veh. Technol.*, vol. 65, no. 8, pp. 6010–6023, Aug. 2016.
- [98] Y. Liu, Z. Ding, M. ElKashlan, and J. Yuan, "Nonorthogonal multiple access in large-scale underlay cognitive radio networks," *IEEE Trans. Veh. Technol.*, vol. 65, no. 12, pp. 10152–10157, Dec. 2016.
- [99] L. Lv, J. Chen, and Q. Ni, "Cooperative non-orthogonal multiple access in cognitive radio," *IEEE Commun. Lett.*, vol. 20, no. 10, pp. 2059–2062, Oct. 2016.
- [100] M. Zeng, G. I. Tsiropoulos, O. A. Dobre, and M. H. Ahmed, "Power allocation for cognitive radio networks employing non-orthogonal multiple access," in *Proc. IEEE Glob. Commun. Conf. (GLOBECOM)*, Washington, DC, USA, Dec. 2016, pp. 1–5.
- [101] N. Zabetian, M. Baghani, and A. Mohammadi, "Rate optimization in NOMA cognitive radio networks," in *Proc. Int. Telecommun. Symp. (ITS)*, Tehran, Iran, Sep. 2016, pp. 62–65.
- [102] Y. Zhang *et al.*, "Energy efficiency optimization in cognitive radio inspired non-orthogonal multiple access," in *Proc. IEEE Int. Symp. Pers. Indoor Mobile Radio Commun. (PIMRC)*, Valencia, Spain, Sep. 2016, pp. 1–6.
- [103] Y. Liu, Z. Qin, M. ElKashlan, Y. Gao, and N. Arumugam, "Non-orthogonal multiple access in massive MIMO aided heterogeneous networks," in *Proc. Glob. Commun. Conf. (GLOBECOM)*, Washington, DC, USA, Dec. 2016, pp. 1–6.
- [104] Y. Liu, Z. Qin, M. ElKashlan, A. Nallanathan, and J. A. McCann, "Non-orthogonal multiple access in large-scale heterogeneous networks," *IEEE J. Sel. Areas Commun.*, to be published.
- [105] F. Fang, H. Zhang, J. Cheng, and V. C. M. Leung, "Energy-efficient resource allocation for downlink non-orthogonal multiple access network," *IEEE Trans. Commun.*, vol. 64, no. 9, pp. 3722–3732, Sep. 2016.
- [106] Y. Zhang, H.-M. Wang, T.-X. Zheng, and Q. Yang, "Energy-efficient transmission design in non-orthogonal multiple access," *IEEE Trans. Veh. Technol.*, vol. 66, no. 3, pp. 2852–2857, Mar. 2017.
- [107] Y. Chen, S. Zhang, S. Xu, and G. Y. Li, "Fundamental trade-offs on green wireless networks," *IEEE Commun. Mag.*, vol. 49, no. 6, pp. 30–37, Jun. 2011.
- [108] S. M. R. Islam, N. Avazov, O. A. Dobre, and K.-S. Kwak, "Power-domain non-orthogonal multiple access (NOMA) in 5G systems: Potentials and challenges," *IEEE Commun. Surveys Tuts.*, vol. 19, no. 2, pp. 721–742, 2nd Quart., 2017.
- [109] R. Hoshyari, F. P. Wathan, and R. Tafazolli, "Novel low-density signature for synchronous CDMA systems over AWGN channel," *IEEE Trans. Signal Process.*, vol. 56, no. 4, pp. 1616–1626, Apr. 2008.
- [110] H. Nikopour and H. Baligh, "Sparse code multiple access," in *Proc. IEEE Int. Symp. Pers. Indoor Mobile Radio Commun. (PIMRC)*, London, U.K., Sep. 2013, pp. 332–336.
- [111] M. Taherzadeh, H. Nikopour, A. Bayesteh, and H. Baligh, "SCMA codebook design," in *Proc. IEEE Veh. Technol. Conf. (VTC Fall)*, Vancouver, BC, Canada, Sep. 2014, pp. 1–5.
- [112] D. Cai, P. Fan, X. Lei, Y. Liu, and D. Chen, "Multi-dimensional SCMA codebook design based on constellation rotation and interleaving," in *Proc. IEEE Veh. Technol. Conf. (VTC Spring)*, Nanjing, China, May 2016, pp. 1–5.
- [113] Y. Wu, S. Zhang, and Y. Chen, "Iterative multiuser receiver in sparse code multiple access systems," in *Proc. Int. Conf. Commun. (ICC)*, London, U.K., Jun. 2015, pp. 2918–2923.
- [114] L. Yang, Y. Liu, and Y. Siu, "Low complexity message passing algorithm for SCMA system," *IEEE Commun. Lett.*, vol. 20, no. 12, pp. 2466–2469, Dec. 2016.
- [115] S. Chen *et al.*, "Pattern division multiple access—A novel nonorthogonal multiple access for fifth-generation radio networks," *IEEE Trans. Veh. Technol.*, vol. 66, no. 4, pp. 3185–3196, Apr. 2017.

- [116] M. A. Naim and J. P. Fonseka, "Selective mapping of coded multi-channel transmission," U.S. Patent 9 667 459, May 30, 2017.
- [117] D. Fang *et al.*, "Lattice partition multiple access: A new method of downlink non-orthogonal multiuser transmissions," in *Proc. Glob. Commun. Conf. (GLOBECOM)*, Washington, DC, USA, Dec. 2016, pp. 1–6.
- [118] J. Zeng, B. Li, X. Su, L. Rong, and R. Xing, "Pattern division multiple access (PDMA) for cellular future radio access," in *Proc. Int. Conf. Wireless Commun. Signal Process. (WCSP)*, Nanjing, China, Oct. 2015, pp. 1–5.
- [119] B. Ren *et al.*, "Advanced IDD receiver for PDMA uplink system," in *Proc. IEEE/CIC Int. Conf. Commun. China (ICCC)*, Chengdu, China, Jul. 2016, pp. 1–6.
- [120] M. A. Naim, J. P. Fonseka, and E. M. Dowling, "A building block approach for designing multilevel coding schemes," *IEEE Commun. Lett.*, vol. 19, no. 1, pp. 2–5, Jan. 2015.
- [121] E. M. Dowling and J. P. Fonseka, "Tiled-building-block trellis encoders," U.S. Patent 8 077 790, Dec. 13, 2011.
- [122] Y.-C. Huang and K. R. Narayanan, "Construction π_A and π_D lattices: Construction, goodness, and decoding algorithms," *IEEE Trans. Inf. Theory*, vol. 63, no. 9, pp. 5718–5733, Sep. 2017.
- [123] P. R. B. da Silva and D. Silva, "Design of lattice network codes based on construction D," in *Proc. Int. Telecommun. Symp. (ITS)*, São Paulo, Brazil, Aug. 2014, pp. 1–5.
- [124] Y. Cai and R. C. de Lamare, "Multistage MIMO receivers based on multi-branch interference cancellation for MIMO-CDMA systems," in *Proc. Int. Symp. Wireless Commun. Syst. (ISWCS)*, Tuscany, Italy, Sep. 2009, pp. 614–618.
- [125] Z. Qin, X. Yue, Y. Liu, Z. Ding, and A. Nallanathan, "A unified non-orthogonal multiple access framework for heterogeneous ultra dense networks," *IEEE Commun. Mag.*, to be published.
- [126] C. Lin and G. Y. Li, "Energy-efficient design of indoor mmWave and sub-THz systems with antenna arrays," *IEEE Trans. Wireless Commun.*, vol. 15, no. 7, pp. 4660–4672, Jul. 2016.
- [127] C. Lin and G. Y. Li, "Terahertz communications: An array-of-subarrays solution," *IEEE Commun. Mag.*, vol. 54, no. 12, pp. 124–131, Dec. 2016.
- [128] Z. Ding, P. Fan, and H. V. Poor, "Random beamforming in millimeter-wave NOMA networks," *IEEE Access*, vol. 5, pp. 7667–7681, 2017.



Yunlong Cai (S'07–M'10–SM'16) received the B.S. degree in computer science from Beijing Jiaotong University, Beijing, China, in 2004, the M.Sc. degree in electronic engineering from the University of Surrey, Guildford, U.K., in 2006, and the Ph.D. degree in electronic engineering from the University of York, York, U.K., in 2010. From 2010 to 2011, he was a Post-Doctoral Fellow with the Electronics and Communications Laboratory, Conservatoire National des Arts et Metiers, Paris, France. Since 2011, he has been with the College of Information Science

and Electronic Engineering, Zhejiang University, Hangzhou, China, where he is currently an Associate Professor. From 2016 to 2017, he was a Visiting Scholar with the School of Electrical and Computer Engineering, Georgia Institute of Technology, Atlanta, GA, USA. His research interests include transceiver design for multiple-antenna systems, sensor array processing, adaptive filtering, full-duplex communications, cooperative and relay communications, and wireless information and energy transfer.



Zhijin Qin (S'13–M'16) received the B.S. degree from the Beijing University of Posts and Telecommunications, China, in 2012, and the Ph.D. degree from the Queen Mary University of London, U.K., in 2016. She was with Imperial College London, U.K., as a Post-Doctoral Research Associate from 2016 to 2017. Since 2017, she has been with the School of Computing and Communications, Lancaster University, U.K., as a Lecturer (Assistant Professor). She was a recipient of the Best Paper Award at Wireless Technology

Symposium 2012.

Her research interests include low-power wide-area network in Internet of Things, compressive sensing and machine learning in wireless communications, and non-orthogonal multiple access for 5G networks. She has served as the Co-Chair of IEEE WoWMoM CORAL Workshop 2017, and a TPC Member for many IEEE conferences, such as GLOBECOM, ICC, and VTC.



Fangyu Cui received the B.S. degree in information and communication engineering from Zhejiang University, Hangzhou, China, in 2014, where he is currently pursuing the Ph.D. degree with the College of Information Science and Electronic Engineering. His research interests include modulation schemes for 5G networks, non-orthogonal multiple access, and physical layer security.



Geoffrey Ye Li (S'93–M'95–SM'97–F'06) received the B.S.E. and M.S.E. degrees from the Department of Wireless Engineering, Nanjing Institute of Technology, Nanjing, China, in 1983 and 1986, respectively, and the Ph.D. degree from the Department of Electrical Engineering, Auburn University, AL, USA, in 1994. He was a Teaching Assistant and then a Lecturer with Southeast University, Nanjing, from 1986 to 1991, a Research and Teaching Assistant with Auburn University from 1991 to 1994, and a Post-Doctoral Research

Associate with the University of Maryland at College Park, MD, USA, from 1994 to 1996. He was with AT&T Labs-Research at Red Bank, NJ, USA, as a Senior and then a Principal Technical Staff Member from 1996 to 2000. Since 2000, he has been with the School of Electrical and Computer Engineering, Georgia Institute of Technology as an Associate Professor and then a Full Professor. He has also been holding a Cheung Kong Scholar title with the University of Electronic Science and Technology of China since 2006.

His general research interests include statistical signal processing and communications, with emphasis on cross-layer optimization for spectral- and energy-efficient networks, cognitive radios and opportunistic spectrum access, and practical issues in LTE systems. He has published around 200 journal papers in addition to over 40 granted patents and numerous conference papers in the above areas. His publications have been cited around 30 000 times and he has been recognized as the World's Most Influential Scientific Mind. He was a recipient of the Highly-Cited Researcher Award by Thomson Reuters, the 2010 Stephen O. Rice Prize Paper Award, the 2017 Award for Advances in Communication from the IEEE Communications Society, the 2013 James Evans Avant Garde Award, the 2014 Jack Neubauer Memorial Award from the IEEE Vehicular Technology Society, and the 2015 Distinguished Faculty Achievement Award from the School of Electrical and Computer Engineering, Georgia Tech. He has been involved in editorial activities for over 20 technical journals for the IEEE, including the Founding Editor-in-Chief of IEEE 5G Tech Focus. He has organized and chaired many international conferences, including the Technical Program Vice-Chair of IEEE ICC'03, the Technical Program Co-Chair of IEEE SPAWC'11, the General Chair of IEEE GlobalSIP'14, and the Technical Program Co-Chair of IEEE VTC'16 (Spring). He was awarded the IEEE Fellow for his contributions to signal processing for wireless communications in 2005.



Julie A. McCann directs the Adaptive Emergent/Embedded/Ephemeral Systems Engineering research group which explores how wireless sensing-based actuator systems interact with physical world and how this knowledge can inform protocols, edge-analytics and delay tolerant control systems. She currently co-directs the Intel Collaborative Research Institute in Urban IoT systems and the cross Imperial Smart Connected Futures Centre who explore IoT, cyber-physical systems, crowdsourcing, and related topics. She is an active programme committee mem-

ber for many of the adaptive computing, sensor network, and communications journals and conferences and has also co-chaired conferences, such as IEEE SECON, ACM Self-Adaptive, and Self-Organising Systems. She has invited to give many keynotes, distinguished seminars and has lead and participated in academic and industrial panel discussions and has been called to advise government on aspects concerning IoT, procurement, and security.