IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received September 22, 2018, accepted October 12, 2018, date of publication October 22, 2018,
date of current version November 14, 2018.

Digital Object Identifier 10.1109/ACCESS.2018.2876598

NOMA-Based Cooperative Opportunistic Multicast Transmission Scheme for Two Multicast Groups: Relay Selection and Performance Analysis

YUFANG ZHANG[1], XIAOXIANG WANG[1], DONGYU WANG[1], YIBO ZHANG[1],
QIANG ZHAO[2], AND QIAN DENG[1]

[1]Key Laboratory of Universal Wireless Communication, Ministry of Education, Beijing University of Posts and Telecommunications, Beijing 100876, China
[2]Space Engineering University, Beijing 101416, China

Corresponding author: Yufang Zhang (yf910309@bupt.edu.cn)

This work was supported in part by the National Natural Science Foundation of China under Grant 61701038 and in part by the Fundamental Research Funds for the Central Universities.

ABSTRACT The orthogonal frequency division multiple access (OFDMA)-based cooperative multicast (CM) technology realizing the intra-cooperation of multicast group (MG) has improved the performance in throughput and coverage. In the MG scenario, users are uniformly distributed. Due to the small effective transmission range of relay user (RU), the cooperation is hindered, which limits the performance of CM. Therefore, non-orthogonal multiple access (NOMA)-based two MGs joint cooperative opportunistic multicast scheme (termed as COM-NOMA) is proposed. In COM-NOMA scheme, two MGs are served as one quasi-MG. First, on the same frequency band, base station (BS) opportunistically multicasts the superposition signal, which is formed by two MGs' signals, to the users with good channel conditions in the two MGs using NOMA way. Second, some successfully receiving users are selected as RUs. The superposition signal is forwarded to the other users by Device-to-Device multicast way. Then, according to the coverage ratio of BS and the decode condition of NOMA, the signal to interference and noise ratio threshold for RU selection is given. Finally, the spectral efficiency of COM-NOMA is analyzed and the approximate expression of system coverage ratio is derived. The simulation results show that, spectral efficiency at unit transmit power (spectral efficiency per watt) and system coverage ratio is improved in comparison with other OFDMA-based CM schemes.

INDEX TERMS Cooperative multicast, coverage ratio, no-orthogonal multiple access, relay selection, signal to interference and noise ratio threshold.

I. INTRODUCTION

Due to the development of multi-media services, mobile data increases dramatically. Providing subscribers with high-quality multimedia service on limited bandwidth becomes a hot topic and has been paid much attention recently in both the industry and the academic. For the multimedia video-centric applications [1] (e.g., video on demand, video games, live video streaming, video conference, video surveillance, etc.) that have high popularity, users request numerously during popular time. To maximize system throughput, operators give preference to multicast [2]. It can be seen that multicast services are still the leading drive in mobile data increase [3]. It means that a large quantity of spectrum resources are occupied by multicast services. The pressure on spectrum resources caused by multicast will become a big issue in future.

Orthogonal frequency division multiple access (OFDMA)-based two-stage cooperative multicast (CM) as an advanced multicast scheme has already been researched in improving system throughput and enhancing coverage.[1] In one time slot, the process of multicast is divided into two stages. In the first stage, through opportunistic multicast technology, base station (BS) multicasts data at high rate so that users with good channel conditions successfully receive (termed as SUs). The other users who are failed to receive are unsuccessful users

[1] The coverage ratio is the ratio of the users who have successfully received among the total MG users

(termed as USUs). In the second stage, some SUs are selected as relay users (RUs) to serve USUs.

In [4], a certain proportion (e.g., 50%) of users in a multicast group (MG) successfully receive information from BS in the first stage, and then each USU receives signals forwarded by all SUs simultaneously. To improve energy efficiency, when USU is in the overlapping coverage area of multiple SUs, it only receives information from these SUs [5]. In [6] and [7], RU is placed on the fixed position in the cell. In order to minimize system average outage probability, the optimal location is found to place RUs in the proposed genie-aided cooperative multicast scheme [6]. A try-best RU selection scheme in [8] chooses the closest SU to USU as relay. Those cooperative multicast schemes implement intra-group collaboration of MG. However, the transmission power of RU is low, and the radius of its effective transmission range (ETR) is far less than the radius of the cell. When MG users are uniformly distributed, SUs sparsely locate in the cell. It is difficult for RU to cover USUs. In the first stage, the BS performs opportunistic multicast through modulation and coding scheme (MCS) selection according to the minimal channel gain of the users with good channel conditions (i.e., SUs). Therefore, after forwarding in the second stage, when the USU's reception SINR is greater than the equivalent signal to interference and noise ratio (SINR) of the first-stage opportunistic multicast, demodulation can be successfully performed. However, in the above studies, SINR threshold for RU selection is equal to average value of the first-stage multicast reception SINR in every time slot, which is set in MCS. It doesn't make sure that USU successfully receives in the instantaneous channel conditions.

Non-orthogonal multiple access (NOMA) as a more efficient multiple access technology in spectral efficiency improvement and user fairness compared with OFDMA, has attracted a lot of attentions [9]. By utilizing the power domain rather than the conventional time and frequency domains, NOMA can significantly improve network throughput [10]. The essential idea of NOMA is that multiple users can share the same frequency resources and use different power levels [11].

NOMA-based cooperative relaying is introduced into multicast in cognitive network. It utilizes the characteristic of NOMA that several signals share the same frequency band to enhance the capacity and fairness. Every receiver can receive all signals from the same frequency band. One unicast user (i.e., the primary user) locates closely to the edge of the cell and its channel condition is poor, while the multicast users (i.e., the secondary user) seat closely to the BS such that they have good channel conditions. In the first stage, unicast user and some multicast users can access the same licensed spectrum through NOMA technology due to the difference in channel conditions between two kinds of users. In the second stage, the multicast users selected as relays forward data to the unicast user [12]–[15]. Even though MG users have already subscribed and paid for the service, as the secondary users, their performance cannot be guaranteed.

All the aforementioned research efforts on multicast are limited into a single MG. A joint transmission scheme based on NOMA to further increase the spectral efficiency, which aims at two MGs, is proposed in [16]. But the scheme serves all MG users. On one hand, its system throughput is still limited by the user with the worst channel condition, which is also the bottleneck problem of multicast and has not been settled. On the other hand, via successive interference cancellation (SIC), the other MG's signal superimposed on the same frequency band with the one MG's signal is cancelled directly. The superposition signal in NOMA can not be fully utilized.

Here, we introduce the NOMA technology for two MGs into the CM. A two-stage COM-NOMA is researched, since its lower signaling overhead and complexity than the multi-stage one. In the first stage, BS sends superposition signal to the users of two MGs who own good channel conditions. In the second stage, the SUs that are selected as RUs directly forward to the remaining users through D2MD. RUs transmit data simultaneously, same as in [6], [17], and [18]. Although the transmission power of the RU is low, the small ETR means small path loss. High reception SINRs can be achieved by USUs. The main contributions of this paper are threefold:

- The NOMA-based cooperative opportunistic multicast scheme serving two MGs on the same frequency band is given. In the first stage, the signals of two MGs are superimposed on the power domain. BS accesses the users with good channel conditions in the two MGs. In the second stage, some SUs are selected to forward the superposition signal for USUs via D2MD technology. Simulation results show that the proposed scheme possesses high power efficiency and enhances coverage ratio with raising number of MG users.
- For the instantaneous channel conditions, according to the number of users in each MG, two methods of achieving SINR threshold for RU selection are given.[2] When the number is small, COM-NOMA scheme makes the first-stage real-time muticast SINR as threshold for USUs to select RUs to ensure they can successfully decode in the second stage. When the number is large, in order to reduce the frequency that BS computes and broadcasts the threshold, the closed-from expression of SINR threshold is derived, according to the extreme value theory and statistical analysis. When the first-stage coverage ratio is known, the corresponding threshold for relay selection can be calculated. Simulation results prove the calculated threshold can satisfy USU's decode condition.
- The approximate closed-form expression of the COM-NOMA coverage ratio is derived. In the process of the second-stage coverage ratio calculation, in order to reduce the complexity of integral calculation, the circular area covered by the RU is approximated to its

[2] Although the choice of method depends on the number of users, the specific number of user boundaries is not the focus of this study.

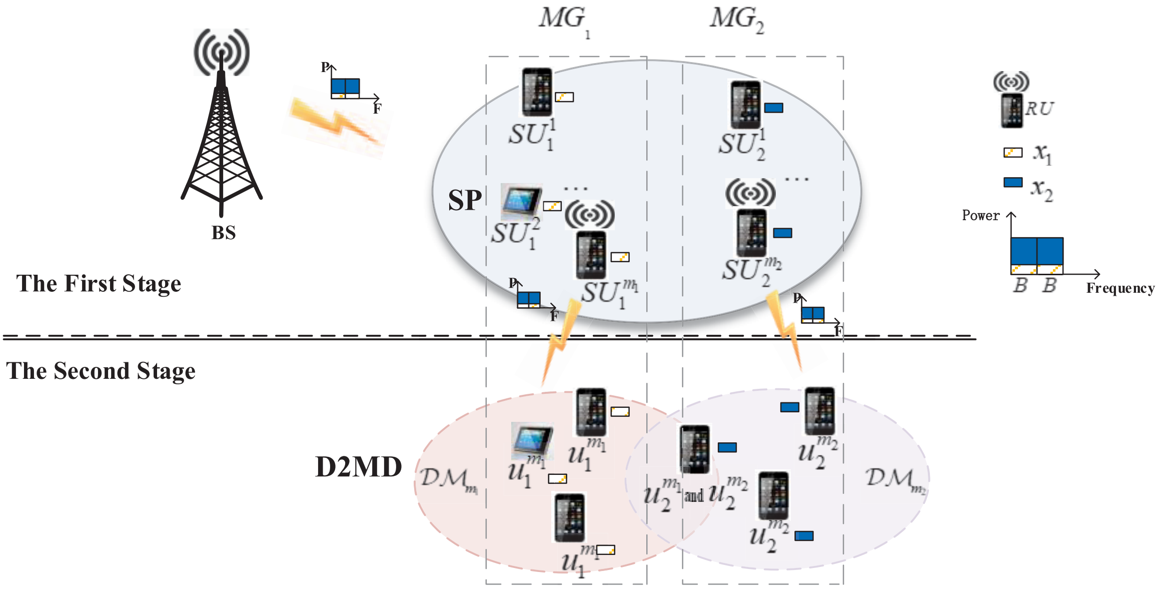


FIGURE 1. System Model of COM-NOMA. In the first stage, BS opportunistically multicasts the superposed signal $X_T = \sqrt{\alpha_1 P}x_1 + \sqrt{\alpha_2 P}x_2$ to the SP which is consisted by the users with good channel conditions in two MGs. In the second stage, the selected RUs forward the superposed signal to its D2MD group $\mathcal{DM}$. The members of $\mathcal{DM}$ come from the two MGs.

circumscribed sector area. Simulation results show that the approximate coverage ratio and the simulated one match well. Futhermore the power allocation factor (PAF), as a key parameter of NOMA, scarcely impacts on the coverage ratio of COM-NOMA scheme.

The rest of this paper is organized as follows. We describe the two-stage cooperative opportunistic multicast based on NOMA in Sec. II. In Sec. III, the second-stage SINR threshold is investigated for USU to select RU. In Sec. IV, we analyze the system coverage ratio and give its approximate close-form expression. The complexity of COM-NOMA scheme is also analyzed in this section. Numerical simulation results and analysis are given in Sec. V. Finally, we conclude this paper and highlight our findings in Sec. VI.

II. SYSTEM MODEL OF COM-NOMA AND SPECTRUM EFFICIENCY ANALYSIS

Consider a downlink two-stage COM-NOMA framework consisting of one BS, two MGs as illustrated in Fig. 1. The two MGs are denoted by MG_i, $i = \{1, 2\}$. The number of users in MG_i is M_i. They are uniformly distributed with a probability density function of distance γ_i from the BS, which is expressed as $f_\gamma(\gamma_i) = \frac{2\gamma_i}{R^2}$, $0 < \gamma_i \le R$, where R is the radius of the cell. They are equipped with one antenna. BS gets all channel state information (CSI). It is assumed that CSI will not change during a transmission time slot T, but be different in different time slots. The channel gains of users in MG_i are in descending order denoted by set $\mathcal{H}_i = \{H_{i,1}, H_{i,2}, \cdots, H_{i,k_i}, \cdots, H_{i,M_i}\}$, $k_i \in [1, M_i]$, where $H_{i,k_i} = \left|h_{i,k_i}\right|^2 \gamma_{i,k_i}^{-\beta}$. h_{i,k_i} is the complex channel coefficient for the link from BS to user k_i and fellows circularly symmetric complex Gaussian distribution $\mathcal{CN} \sim (0, 1)$. γ_{i,k_i} denotes the Euclidean distance between them. β is the path loss parameter. Time slot is divided into two stages, i.e., T_1 and T_2, $T_1 + T_2 = T$. The work in [19] has proved that the maximal system throughput can be achived, when $T_1 = T_2 = T/2$. It is also assumed in this paper. Downlinks from BS to every users and between any two users are subject to Rayleigh fading, propagation path loss and additive white Gaussian noise (AWGN). COM-NOMA still occupies the bandwidth originally allocated to the two MGs in the OFDMA system.

In the first stage, through opportunistic multicast BS only serves a fraction of MG users who have good channel conditions. In the transmission of BS, two MGs' signals are transmitted simultaneously by NOMA signaling on the same frequency band. The remaining users with bad channel conditions give up receiving the signal that leads to failed decoding. In the second stage, BS mutes and some of the SUs are selected as RUs, as illustrated in Fig.1. It is assumed that all users are accommodating. They are willing to forward data for other users.

A. THE FIRST STAGE TRANSMISSION

In the first stage of COM-NOMA, BS opportunistically multicasts to the users with good channel conditions of two MGs at coverage ratio C_1 using the joint transmission scheme proposed in [16]. They are denoted by set $\mathcal{S}_i$, when they belong to MG_i. $\mathcal{S}_i = \left\{SU_{i,1}, SU_{i,2}, \cdots, SU_{i,m_i}, \cdots, SU_{i,s_i}\right\}$, $s_i = C_1 M_i$, $m_i \in [1, s_i]$. The corresponding channel gain set is $\mathcal{H}_{\mathcal{S}_i} = \left\{H_{i,1}, H_{i,2}, \cdots, H_{i,m_i}, \cdots, H_{i,s_i}\right\}$. Without loss of generality, $H_{1,s_1} > H_{2,s_2}$ is assumed here. $\mathcal{H}_{\mathcal{S}_i}$ is divided into three subgroups, $\mathcal{H}_{\mathcal{S}_{i,b}}$, $\mathcal{H}_{\mathcal{S}_{i,m}}$ and $\mathcal{H}_{\mathcal{S}_{i,w}}$, by Maximum

Algorithm 1 Maximum and Minimum Group Method

```
Require: Two MPs' channel gains sets:$\mathcal{H}_{\mathcal{S}_i}, i = \{1, 2\}$
Ensure: $\mathcal{H}_{\mathcal{S}_i,b} \cup \mathcal{H}_{\mathcal{S}_i,m} \cup \mathcal{H}_{\mathcal{S}_i,w} = \mathcal{H}_{\mathcal{S}_i}$,
        $\mathcal{H}_{\mathcal{S}_i,b} \cap \mathcal{H}_{\mathcal{S}_i,m} \cap \mathcal{H}_{\mathcal{S}_i,w} = \varnothing$
if $\exists H_{1,m_1} < H_{2,s_2}$ then
    $\mathcal{H}_{\mathcal{S}_1,w} = \{H_{1,m_1} | H_{1,m_1} < H_{2,s_2}\}$ and $\mathcal{H}_{\mathcal{S}_2,w} = \varnothing$;
else if $\exists H_{1,m_1} = H_{2,s_2}$ then
    $\mathcal{H}_{\mathcal{S}_1,w} = \{H_{1,m_1} | H_{1,m_1} = H_{2,s_2}\}$ and $\mathcal{H}_{\mathcal{S}_2,w} = \{H_{2,s_2}\}$;
else
    $\mathcal{H}_{\mathcal{S}_2,w} = \{H_{2,m_2} | H_{2,m_2} < H_{1,s_1}\}$ and $\mathcal{H}_{\mathcal{S}_1,w} = \varnothing$;
end if
if $\exists H_{1,m_1} > H_{2,1}$ then
    $\mathcal{H}_{\mathcal{S}_1,b} = \{H_{1,m_1} | H_{1,m_1} > H_{2,s_2}\}$ and $\mathcal{H}_{\mathcal{S}_2,b} = \varnothing$;
else if $\exists H_{1,m_1} = H_{2,1}$ then
    $\mathcal{H}_{\mathcal{S}_1,b} = \{H_{1,m_1} | H_{1,m_1} = H_{2,s_2}\}$and $\mathcal{H}_{\mathcal{S}_2,b} = \{H_{2,1}\}$;
else
    $\mathcal{H}_{\mathcal{S}_2,b} = \{H_{2,m_2} | H_{2,m_2} > H_{1,1}\}$ and $\mathcal{H}_{\mathcal{S}_1,b} = \varnothing$;
end if
```

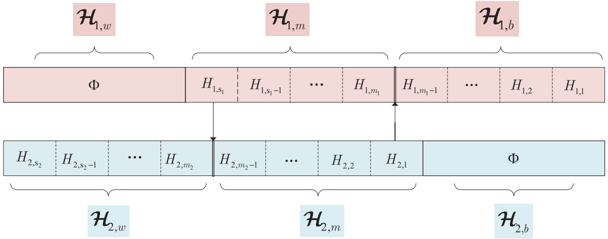


FIGURE 2. **Maximum and Minimum group method in the case of $H_{1,s_1} > H_{2,s_2}$ *and* $H_{1,1} > H_{2,1}$.**

and Minimum group method (MMGM) as is shown in **Algorithm 1**. And $\mathcal{H}_{\mathcal{S}_i,b} > \mathcal{H}_{\mathcal{S}_i,m} > \mathcal{H}_{\mathcal{S}_i,w}$. In Fig.2, the case of $H_{1,s_1} > H_{2,s_2}$ *and* $H_{1,1} > H_{2,1}$ is shown. The corresponding user subsets are $\mathcal{S}_i^b$, $\mathcal{S}_i^m$ and $\mathcal{S}_i^w$.

Because $\mathcal{H}_{\mathcal{S}_2,w} < \mathcal{H}_{\mathcal{S}_1}$, BS transmits the superposition signal in NOMA to their corresponding user subgroup pair (SP), $\mathcal{S}_2^w$ and $\mathcal{S}_1$ ($\mathcal{S}_1 = \mathcal{S}_1^m \cup \mathcal{S}_1^b$). And the transmission rate is decided by the least channel gain, H_{2,s_2}. The transmission signal is $X_T = \sqrt{\alpha_1 P} x_1 + \sqrt{\alpha_2 P} x_2$, where P is the transmission power of BS. α_1 and α_2 are the PAFs of the paired subgroups $\mathcal{S}_1$ and $\mathcal{S}_2^w$, $\alpha_1 + \alpha_2 = 1$. Since $\mathcal{H}_{\mathcal{S}_2,w} < \mathcal{H}_{\mathcal{S}_1}$, $0 < \alpha_1 < \alpha_2 < 1$. x_i is the demand signal of MG_i. The received signals of SU_{1,s_1} and SU_{2,s_2} are given by

$$Y_{1,s_1} = h_{1,s_1}(\sqrt{\alpha_1 P \gamma_{1,s_1}^{-\beta}} x_1 + \sqrt{\alpha_2 P \gamma_{1,s_1}^{-\beta}} x_2) + \omega_n, \quad (1)$$

$$Y_{2,s_2} = h_{2,s_2}(\sqrt{\alpha_1 P \gamma_{2,s_2}^{-\beta}} x_1 + \sqrt{\alpha_2 P \gamma_{2,s_2}^{-\beta}} x_2) + \omega_n, \quad (2)$$

where ω_n is AWGN with variance δ_n^2.

In NOMA, the signal assigned with large PAF is decoded directly. x_2 is successfully decoded by the SUs in $\mathcal{S}_2^w$. Due to the channel gains in $\mathcal{H}_{2,b}$ and $\mathcal{H}_{2,m}$ are lager than H_{2,s_2}, x_2 can also be decoded from the superposition signal by users in $\mathcal{S}_2^m$ and $\mathcal{S}_2^b$. Multicast rate is decided by the member who owns the worst channel condition. The equivalent reception SINR of SUs in $\mathcal{S}_2$ is

$$SINR_{\mathcal{S}_2} = \frac{\alpha_2 P \left|h_{2,s_2}\right|^2 \gamma_{2,s_2}^{-\beta}}{\alpha_1 P \left|h_{2,s_2}\right|^2 \gamma_{2,s_2}^{-\beta} + 2N_0}, \quad (3)$$

where $N_0 = B\delta_n^2$, B is the bandwidth allocated to one MG in the OFDMA system. Because COM-NOMA occupies the bandwidth allocated to the two MGs, the power of AWGN is $2N_0$.

As the interference signal of x_1, the equivalent reception SINR of signal x_2 received by the SUs in $\mathcal{S}_1$ is

$$SINR_{\mathcal{S}_1}^{x_2} = \frac{\alpha_2 P \left|h_{1,s_1}\right|^2 \gamma_{1,s_1}^{-\beta}}{\alpha_1 P \left|h_{1,s_1}\right|^2 \gamma_{1,s_1}^{-\beta} + 2N_0}. \quad (4)$$

Given $H_{1,s_1} > H_{2,s_2}$, $SINR_{\mathcal{S}_1}^{x_2} > SINR_{\mathcal{S}_2}$. So they can decode the interference signal x_2 firstly. After SIC, x_2 is cancelled and x_1 is obtained. Thanks to their larger SINRs than $SINR_{\mathcal{S}_2}$, SUs in $\mathcal{S}_1$ decode their demand signal x_1 successfully. The equivalent reception SINR of x_1[3] after SIC is

$$SINR_{\mathcal{S}_1} = \frac{\alpha_1 P \left|h_{1,s_1}\right|^2 \gamma_{1,s_1}^{-\beta}}{2N_0}. \quad (5)$$

Following the analysis in [16], the throughput of the opportunistic multicast in the first stage is

$$R_1 = 2B \sum_i C_1 M_i \log_2 \left(1 + SINR_{\mathcal{S}_i}\right). \quad (6)$$

In summery, MG_1 and MG_2 are regarded as one quasi-MG in the first stage. BS transmits the signal X_T to the quasi-MG by opportunistic multicast with fixed coverage ratio C_1. After the first stage, SU_{i,m_i} out of $\mathcal{S}_i$ who has already simultaneously received x_1 and x_2, can be enlisted as candidate relays for USUs.

B. THE SECOND STAGE TRANSMISSION

In the second stage, the selected RU forwards the first-stage reception signals in Device-to-Device multicast (D2MD) [20] to USU. The set of remaining USUs is denoted by $\mathcal{U}$.

One important fact in D2D communication is that the radius of ETR (e.g., 100m) [21], [22] is much smaller than that of cell due to the low transmission power of the RU. Therefore, only SUs in the vicinity of a USU (that is, SUs in the USU's ETR) can serve as candidate relays. However, the small radius of ETR also means USU has probability to get higher SINR than first-stage SUs so that the demand signal can be decoded from the superposition signal. The radius of ETR is denoted by R_0, $R_0 << R$.

[3] The interference signal has already been cancelled. SINR here is to simplify writing of (5).

From the analysis of previous subsection, the larger received SINR than $SINR_{S_2}$ guarantees successfully decoding. Therefore, the SINR threshold σ_0 for RU selection is not smaller than the minimum SINR in the first-stage opportunistic multicast, i.e., $\sigma_0 \geq SINR_{\mathcal{S}_2}$. The RU set of u is denoted by $\mathcal{R}_u = \{SU_{i,m_i} | \gamma_{m_i,u} <= R_0, SINR_{m_i,u}^{x_2} >= \sigma_0\}$. $\gamma_{m_i,u}$ is the distance from u to its RU SU_{i,m_i}. $SINR_{m_i,u}^{x_2}$ denotes the reception SINR of u from SU_{i,m_i} for signal x_2.

The reception signal of u from SU_{i,m_i} is

$$Y_u = h_{m_i,u}(\sqrt{\alpha_1 P_D} x_1 + \sqrt{\alpha_2 P_D} x_2) + \omega_n. \tag{7}$$

P_D is RU's transmission power, which is much smaller than P. The reception SINR is

$$SINR_{m_i,u}^{x_2} = \frac{\alpha_2 P_D \left|h_{m_i,u}\right|^2 \gamma_{m_i,u}^{-\beta}}{\alpha_1 P_D \left|h_{m_i,u}\right|^2 \gamma_{m_i,u}^{-\beta} + 2N_0}. \tag{8}$$

The USUs served by the same relay SU_{i,m_i} consist a D2MD group, which is denoted by $\mathcal{DM}_{m_i}$. The number of USUs in $\mathcal{DM}_{m_i}$ is $|\mathcal{DM}_{m_i}|$. The set of the D2MD transmitters is denoted by $\mathcal{R}_{D2MD}$, $\mathcal{R}_{D2MD} = \bigcup\limits_{u \in \mathcal{U}} \mathcal{R}_u$. The RU selection and transmission strategy is shown in **Algorithm 2**.

Algorithm 2 RU Selection and Tranmission Strategy

```
1: BS broadcasts a training signal containing a quantized version of $\sigma_0$ to the users in the two MGs.
2: Every MG user broadcasts a flag message, Success or Failure according to it is a SU or not. If the flag message is Failure, the $SU_{i,m_i}$, measures the channel gain between the USU $u$ and itself, which is denoted as $H_{m_i,u}$. As thus, the CSI of other MG users within the ETR are obtained.
3: $SU_{i,m_i}$ detects obtained channel gain of every USU. If there are channel gians of USUs making the value of (8) larger than $\sigma_0$. This $SU_{i,m_i}$ is selected as a RU. Those USUs who make the value of (8) larger than $\sigma_0$ are the D2MD members served by $SU_{i,m_i}$.
4: Every selected RU forwards reception signal in D2MD.
```

For signal x_2, the equivalent reception SINR in $\mathcal{DM}_{m_i}$ is

$$SINR_{\mathcal{DM}_{m_i}} = \min_{u \in \mathcal{DM}_{m_i}} SINR_{m_i,u}^{x_2}. \tag{9}$$

$\mathcal{U}_i^{m_i}$ is the collection of USUs out of $\mathcal{DM}_{m_i}$. Its subscript i represents the MG that the USU belongs to. Its superscript m_i means the transmitter of $\mathcal{DM}_{m_i}$, i.e., SU_{i,m_i}. $|\mathcal{U}_i^{m_i}|$ represents the number of USUs in $\mathcal{U}_i^{m_i}$. $u_i^{m_i}$ is the USU in $\mathcal{U}_i^{m_i}$. The D2MD reception SINRs of $u_i^{m_i}$ are

$$SINR_1^{m_i} = \min_{u \in \mathcal{U}_1^{m_i}} \left(\frac{\alpha_1 P_D \left|h_{m_i,u}\right|^2 \gamma_{m_i,u}^{-\beta}}{2N_0} \right), \tag{10}$$

$$SINR_2^{m_i} = SINR_{\mathcal{DM}_{m_i}}. \tag{11}$$

It is assumed that the cyclic prefix (CP) is longer than the maximum delay of the equivalent multipath channel, all the received signals arriving within the CP duration can be added up to construct a stronger signal, which is denoted as CP combining [6]. So the throughput of the second stage is

$$R_2 = 2B \sum_{SU_{i,m_i} \in \mathcal{R}_{D2MD}} \sum_i \left|\mathcal{U}_i^{m_i}\right| \log_2\left(1 + SINR_i^{m_i}\right). \tag{12}$$

The spectral efficiency of COM-NOMA is

$$E = \frac{1}{2} \frac{(R_1 + R_2)}{2B}. \tag{13}$$

III. ANALYSIS OF SINR THRESHOLD FOR RU SELECTION

In this section, we analyze the SINR threshold σ_0. In the second stage, if SU_{i,m_i} is a RU of USU u, the distance between them should be less than R_0, and he (she) can provide USU u with greater SINR than $SINR_{\mathcal{S}_2}$, which is decided by the opportunistic multicast in the first stage. From the analysis in Sec. II-A, based on the instantaneous channel conditions, when the number of MG users is small, BS is able to compute and broadcast the threshold frequently, and $\sigma_0 = SINR_{\mathcal{S}_2}$. However, when the number of MG users is large, it is a heavy load for BS. To reduce the burden of the BS and get effective SINR threshold, statistical property of $SINR_{\mathcal{S}_2}$ is taken into account.

According to the extreme value theory and lemma [23], given coverage ratio C_1, the transmission rate per unit bandwidth of opportunistic multicast R_{2,s_2}, which is decided by the SU with the smallest channel gain in the first stage, follows normal distribution. $R_{2,s_2} \sim \mathcal{CN}\left(F_{R_{2,s_2}}^{-1}(1 - C_1), \frac{\delta^2(C_1)}{N}\right)$, $\delta^2(C_1) = \frac{C_1(1-C_1)}{\left[f_{R_{2,s_2}}\left(F_{R_{2,s_2}}^{-1}(1-C_1)\right)\right]^2}$, where N is the total number of users, $N = M_1 + M_2$, and $R_{2,s_2} = \log_2(1 + SINR_{\mathcal{S}_2})$. So the cumulative distribution function (CDF) and probability distribution function (PDF) of R_{2,s_2}, $F_{R_{2,s_2}}$ and $f_{R_{2,s_2}}$ are the keys to get the its distribution.

The multicast reception SINR of users in $\mathcal{S}_2$ in the first stage is expressed by

$$SINR_{\mathcal{S}_2} = \frac{\alpha_2 \frac{P}{N_0} \left|h_{2,s_2}\right|^2 {\gamma_{2,s_2}}^{-\beta}}{\alpha_1 \frac{P}{N_0} \left|h_{2,s_2}\right|^2 {\gamma_{2,s_2}}^{-\beta} + 2}, \tag{14}$$

with $x = \frac{P}{N_0} \left|h_{2,s_2}\right|^2 \gamma_{2,s_2}^{-\beta}$ and $\rho_0 = \frac{P}{N_0}$. Since $h_{2,s_2} \sim \mathcal{CN}(0, 1)$, for any given γ_{2,s_2}, x follows exponential distribution with parameter $\lambda_{\mathcal{S}_2} = \frac{{\gamma_{2,s_2}}^{\beta}}{\rho_0}$. The CDF of X is

$$F_{X|\gamma_{2,s_2}}(x) = 1 - \exp(-\lambda_{\mathcal{S}_2} x). \tag{15}$$

So in COM-NOMA, let $y = \frac{\alpha_2 x}{\alpha_1 x + 2}$, $SINR_{\mathcal{S}_2} = y$. The conditional distribution of $Y = \frac{\alpha_2 X}{\alpha_1 X + 2}$ is

$$\begin{aligned} F_{Y|\gamma_{2,s_2}}(y) &= P(Y \leq y) \\ &= P\left(X \leq \frac{2Y}{(\alpha_2 - \alpha_1 Y)}\right) \\ &= 1 - exp\left(-\lambda_{\mathcal{S}_2} \frac{2y}{(\alpha_2 - \alpha_1 y)}\right), \end{aligned} \tag{16}$$

and the marginal CDF of Y is

$$\begin{aligned}&F_Y\left(y\right)\\&=\int_0^R F_{Y|\gamma_{2,s_2}}\left(y\right)f_\gamma\left(\gamma_{2,s_2}\right)d\gamma_{2,s_2}\\&=\int_0^R\left[1-exp(-\frac{N_0(\gamma_{2,s_2})^\beta}{P}\frac{2y}{(\alpha_2-\alpha_1 y)})\right]*\frac{2\gamma_{2,s_2}}{R^2}d\gamma_{2,s_2}\\&=1-\frac{1}{\beta R^2}\left[\frac{(\alpha_2-\alpha_1 y)\rho_0}{2y}\right]^{\frac{2}{\beta}}*\Gamma(\frac{2}{\beta},\frac{2yR^\beta}{(\alpha_2-\alpha_1 y)\rho_0}),\end{aligned}\tag{17}$$

where $\Gamma(a,x)=\int_0^x t^{a-1}\exp(-t)dt$ is the incomplete gamma function.

According to Shannon theory, $y=2^{r_{2,s_2}}-1$. The CDF of R_{2,s_2} is

$$\begin{aligned}F_{R_{2,s_2}}\left(r_{2,s_2}\right)&=F_Y\left(2^{r_{2,s_2}/(2B)}-1\right)\\&=1-\frac{1}{\beta R^2}\left(\frac{\rho_0}{\nu_2}\right)^{\frac{2}{\beta}}\Gamma\left(\frac{2}{\beta},\frac{\nu_2 R^\beta}{\rho_0}\right),\end{aligned}\tag{18}$$

where $\nu_2=\frac{2\left(2^{r_{2,s_2}}-1\right)}{\alpha_2-\alpha_1\left(2^{r_{2,s_2}}-1\right)}$. So the PDF of R_{2,s_2} is

$$\begin{aligned}&f_{R_{2,s_2}}\left(r_{2,s_2}\right)\\&=\frac{dF_{R_{2,s_2}}\left(r_{2,s_2}\right)}{dr_{2,s_2}}\\&=\frac{d\int_0^R\left[1-exp\left(-\frac{N_0\left(\gamma_{2,s_2}\right)^\beta}{P}\nu_2\right)\right]*\frac{2\gamma_{2,s_2}}{R^2}d\gamma_{2,s_2}}{dr_{2,s_2}}\\&=\int_0^R\underbrace{\frac{d\left\{\left[1-exp\left(-\frac{N_0\left(\gamma_{2,s_2}\right)^\beta}{P}\nu_2\right)\right]*\frac{2\gamma_{2,s_2}}{R^2}\right\}}{dr_{2,s_2}}}_{\mathcal{O}}d\gamma_{2,s_2}\\&=\frac{2\nu_1}{\beta\rho_0 R^2}\left(\frac{\rho_0}{\nu_2}\right)^{\frac{2}{\beta}+1}\Gamma\left(\frac{2}{\beta}+1,\frac{\nu_2}{\rho_0}R^\beta\right).\end{aligned}\tag{19}$$

And

$$\mathcal{O}=\frac{N_0\left(\gamma_{2,s_2}\right)^\beta}{P}\nu_1 exp\left(-\frac{N_0\left(\gamma_{2,s_2}\right)^\beta}{P}\nu_2\right)*\frac{2\gamma_{2,s_2}}{R^2},\tag{20}$$

$$\nu_1=\frac{a_2\cdot 2^{1+r_{2,s_2}/(2B)}\ln 2}{\left(\alpha_2-\alpha_1\left(2^{r_{2,s_2}/(2B)}-1\right)\right)^2}.$$

From (18) and (19), given a first-stage coverage ratio C_1, the mean value and variance of minimal reception rate per unit bandwidth R_{2,s_2}, that supports SUs to decode their demand signals can be obtained. According to thrice standard error principle of normal distribution, the rate threshold for relay selection in large number of MG users is set to

$$R_{\sigma_0}\left(C_1\right)=F^{-1}_{R_{2,s_2}}\left(1-C_1\right)+3\cdot\frac{\delta\left(C_1\right)}{\sqrt{N}}.\tag{21}$$

For the instantaneous channel conditions, the event that the first-stage multicast rate R_{2,s_2} is greater than R_{σ_0} can be considered as a small probability event. When the rate threshold is set to R_{σ_0}, it can be ensured that the reception rate of USU in D2MD group is greater than that of the first stage, so that the USU is able to decode the demand data successfully from the NOMA signaling.

Because the closed-form expression of $F^{-1}_{R_{2,s_2}}$ is difficult to be derived, to obtain further insights of $R_{\sigma_0}\left(C_1\right)$, its approximate expression is given, when $\frac{P}{N_0}\to\infty$.

$$\begin{aligned}&F^A_{R_{2,s_2}}\left(r_{2,s_2}\right)\\&\quad=\int_0^R\left[1-exp(-\frac{N_0(\gamma_{2,s_2})^\beta}{P}\frac{2\left(2^{r_{2,s_2}}-1\right)}{\alpha_2-\alpha_1\left(2^{r_{2,s_2}}-1\right)})\right]\\&\qquad\cdot\frac{2\gamma_{2,s_2}}{R^2}d\gamma_{2,s_2}\\&\quad\approx\int_0^R\frac{N_0}{P}\frac{2\left(2^{r_{2,s_2}}-1\right)}{\alpha_2-\alpha_1\left(2^{r_{2,s_2}/(2B)}-1\right)}\frac{2(\gamma_{2,s_2})^{\beta+1}}{R^2}d\gamma_{2,s_2}\\&\quad=\frac{N_0}{P}\frac{R^\beta}{\beta+2}\cdot\frac{4\left(2^{r_{2,s_2}}-1\right)}{\alpha_2-\alpha_1\left(2^{r_{2,s_2}}-1\right)}.\end{aligned}\tag{22}$$

The first order Taylor series approximation $exp(-t)=1-t$ for t close to 0 is used here. The approximate expression of $f_{R_{2,s_2}}\left(r_{2,s_2}\right)$ is

$$f^A_{R_{2,s_2}}\left(r_{2,s_2}\right)=\frac{dF^A_{R_{2,s_2}}\left(r_{2,s_2}\right)}{dr_{2,s_2}}=\frac{N_0}{P}\frac{R^\beta}{\beta+2}\cdot 4\nu_1.\tag{23}$$

After simple algebraic operation, the mean and standard deviation of R_{2,s_2} are $\log_2\left(\frac{\nu_3+(1-C_1)}{\nu_3+a_1(1-C_1)}\right)$ and $\frac{\nu_3 a_2\sqrt{C_1(1-C_1)/N}}{ln2\cdot(\nu_3+(1-C_1))(\nu_3+a_1(1-C_1))}$. So

$$\begin{aligned}R_{\sigma_0}\left(C_1\right)=&\log_2\left(\frac{\nu_3+\left(1-C_1\right)}{\nu_3+a_1\left(1-C_1\right)}\right)\\&+3\cdot\frac{\nu_3 a_2\sqrt{C_1\left(1-C_1\right)/N}}{ln2\cdot\left(\nu_3+\left(1-C_1\right)\right)\left(\nu_3+a_1\left(1-C_1\right)\right)},\end{aligned}\tag{24}$$

where $\nu_3=\frac{4N_0}{P}\frac{R^\beta}{\beta+2}$.

Substituting (24) into $\sigma_0=2^{R_{\sigma_0}(C_1)}-1$, SINR threshold σ_0 for large number of MG users is got here.

In summary, the SINR threshold σ_0 is given for RU selection in different number of users in each MG M_i. When M_i is small, $\sigma_0=SINR_{\mathcal{S}_2}$. When M_i is large, $\sigma_0=2^{R_{\sigma_0}(C_1)}-1$, which is only decided by the coverage ratio in the first stage.

IV. PERFORMANCE ANALYSIS

In this section, the coverage ratio and complexity of the proposed COM-NOMA scheme are analyzed. The approximate close-form expression of coverage ratio is derived.

A. ANALYSIS OF COVERAGE RATIO PERFORMANCE

SUs are successfully served in the first stage. To ensure that remaining USUs can decode, some SUs are selected as RUs to forward the superposition signal in the second stage. The coverage ratio of COM-NOMA system is expressed by

$$C_{COM-NOMA} = C_1 + C_2, \tag{25}$$

where C_2 is the coverage ratio in the second stage.

For USU u, if he (she) owns at least one RU, according to the selection scheme, every RU locates within the ETR of u, and provides its D2MD members with larger SINR than the threshold σ_0.

In NOMA technology, the two paired users' demand signals are superposed on the same frequency band based on difference in their channel conditions. At receiving end, the signal demanded by the user with bad channel condition is decoded firstly, which is x_2 in this paper. If users in whichever MG decode x_2 successfully, they can decode their demand signal directly or after SIC. So the coverage ratio of the second stage is equal to the probability that USU successfully decodes x_2 from one of his (her) RUs.

For x_2, the reception SINR of u from its RU SU_{i,m_i} is

$$\begin{aligned} SINR_{m_i,u} &= \frac{\alpha_2 P_D \left|h_{m_i,u}\right|^2 {\gamma_{m_i,u}}^{-\beta}}{\alpha_1 P_D \left|h_{m_i,u}\right|^2 {\gamma_{m_i,u}}^{-\beta} + 2N_0} \\ &= \frac{\alpha_2 \frac{P_D}{N_0} \left|h_{m_i,u}\right|^2 {\gamma_{m_i,u}}^{-\beta}}{\alpha_1 \frac{P_D}{N_0} \left|h_{m_i,u}\right|^2 {\gamma_{m_i,u}}^{-\beta} + 2}, \end{aligned} \tag{25}$$

with $x_{m_i,u} = \frac{P_D}{N_0}\left|h_{m_i,u}\right|^2 {\gamma_{m_i,u}}^{-\beta}$, $\rho_1 = \frac{P_D}{N_0}$. Let $y_{m_i,u} = \frac{\alpha_2 x_{m_i,u}}{\alpha_1 x_{m_i,u}+2}$. CDF of the instantaneous SNR $x_{m_i,u}$ is given by

$$F_{X_{m_i,u}}\left(x_{m_i,u}\right) = 1 - \exp\left(-\lambda_{m_i,u} x_{m_i,u}\right), \tag{26}$$

with exponential distribution parameter $\lambda_{m_i,u} = \frac{(\gamma_{m_i,u})^{\beta}}{\rho_1}$. So the conditional CDF of $Y_{m_i,u} = \frac{\alpha_2 X_{m_i,u}}{\alpha_1 X_{m_i,u}+2}$ is

$$F_{Y_{m_i,u}|\gamma_{m_i,u}}\left(y_{m_i,u}|\gamma_{m_i,u}\right) = 1-exp\left(-\frac{{\gamma_{m_i,u}}^{\beta}}{\rho_1}\frac{2y_{m_i,u}}{\alpha_2 - \alpha_1 y_{m_i,u}}\right). \tag{27}$$

The probability that the SINR provided by the selected RU is higher than σ_0 is

$$P(SINR_{m_i,u} > \sigma_0|\gamma_{m_i,u}) = 1 - F_{Y_{m_i,u}|\gamma_{m_i,u}}\left(\sigma_0|\gamma_{m_i,u}\right). \tag{28}$$

At the same time, SU_{i,m_i} must be within the ETR with radius R_0 of USU u. Due to $R_0 << R$, the circle transmission region where SU_{i,m_i} may exist can be approximated to its circumscribed sector field, as is shown in Fig.3. The probability that there is a SU being selected as u's RU is

$$\begin{aligned} &P_{m_i}(\gamma_u) \\ &\quad = \theta(\gamma_u)\int_{\gamma_u - R_0}^{\gamma_u + R_0} P(SINR_{m_i,u} > \sigma_0|\gamma_{m_i,u})\cdot f_{\gamma}\left(\gamma_{m_i,u}\right) d\gamma_{m_i,u}, \end{aligned} \tag{29}$$

where $\theta(\gamma_u) = arcsin(\frac{R_0}{\gamma_u})$ and γ_u is the distance between u and BS. Because of the uniform distribution of MG users in the cell, the joint PDF of the distance from the SU_{i,m_i} to u is expressed by $f_{\gamma}\left(\gamma_{m_i,u}\right) = \frac{2\gamma_{m_i,u}}{R^2}$.

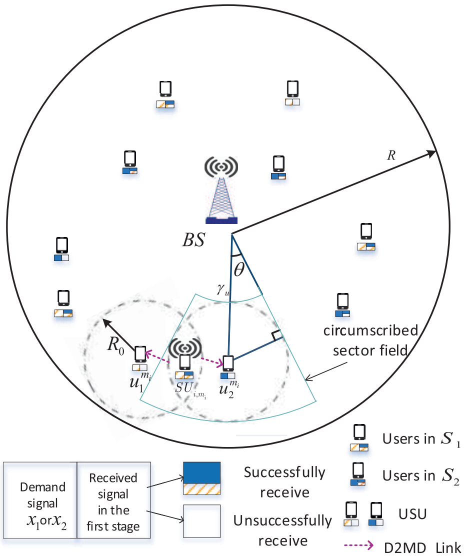


FIGURE 3. Illustration of the second step.

Aiming at any one USU, it is assumed that the distance from BS is larger than R_0. The average probability that USU is served by one RU is

$$P_{m_i} = \frac{1}{2\pi}\int_{R_0}^{R} P_{m_i}(\gamma_u) f_{\gamma}\left(\gamma_u\right) d\gamma_u. \tag{30}$$

Because a USU can be served by at least one RU, the probability that the USU can successfully receive data from his (her) RUs is

$$P_u = 1 - (1 - P_{m_i})^{(M_1+M_2)C_1}. \tag{31}$$

The coverage ratio of the second stage

$$C_2 = (1 - C_1)P_u. \tag{32}$$

From the analysis above, it is hard to find the close-form expression of P_{m_i}. We obtain its asymptotically tight approximation in (36), as shown at the bottom of the next page, when $\rho_1 \to \infty$. Where

$$\begin{aligned} P(SINR_{m_i,u} > \sigma_0|\gamma_{m_i,u}) &= exp\left(-\frac{(\gamma_{m_i,u})^{\beta}}{\rho_1}\frac{2\sigma_0}{(\alpha_2 - \alpha_1\sigma_0)}\right) \\ &\approx 1 - \frac{(\gamma_{m_i,u})^{\beta}}{\rho_1}\frac{2\sigma_0}{(\alpha_2 - \alpha_1\sigma_0)} \approx 1. \end{aligned} \tag{33}$$

Substituting (36) into (32), after simple algebraic operation, C_2 is derived. The approximate $C^A_{COM-NOMA}$ is expressed by.

$$C^A_{COM-NOMA} \approx C_1 + (1-C_1)\left(1-(1-P^A_{m_i})^{C_1(M_1+M_2)}\right). \tag{34}$$

B. ANALYSIS OF COMPLEXITY

The complexity of the proposed COM-NOMA scheme is analyzed in the following. In the transmission process, there are three main steps affecting the complexity. The first step: the channel gains of users in the two multicast groups MG_1 and MG_2, are sorted in descending order, separately. The complexity is $\mathcal{O}(M_1 \log_2 M_1) + \mathcal{O}(M_2 \log_2 M_2)$. The second step: through MMGM, $\mathcal{H}_{\mathcal{S}_i}$ is divided into three subgroups, $\mathcal{H}_{\mathcal{S}_i,b}$, $\mathcal{H}_{\mathcal{S}_i,m}$ and $\mathcal{H}_{\mathcal{S}_i,w}$. According to MMGM, it takes up to M_1 comparisons to divide $\mathcal{H}_{\mathcal{S}_i}$ into three subgroups. The complexity is equal to $\mathcal{O}(C_1 M_1)$. The third step: in RU selection, there are $C_1(M_1+M_2)$ SUs (they are candidate RUs) and $(1-C_1)(M_1+M_2)$ USUs. The RU set of USU u is $\mathcal{R}_u = \{SU_{i,m_i}|\gamma_{m_i,u} <= R_0, SINR^{x_2}_{m_i,u} >= \sigma_0\}$. So the complexity is $\mathcal{O}((C_1)(1-C_1)(M_1+M_2)^2)$. In summary, system complexity is the sum of the complexity of these three steps.

$$\begin{aligned}\mathcal{O}(M_1 \log_2 M_1) &+ \mathcal{O}(M_2 \log_2 M_2) + \mathcal{O}(C_1 M_1)\\ &+ \mathcal{O}(C_1(1-C_1)(M_1+M_2)^2))\\ &= \mathcal{O}(C_1(1-C_1)(M_1+M_2)^2).\end{aligned} \tag{35}$$

V. PERFORMANCE EVALUATION

To illustrate the performance of the proposed COM-NOMA scheme and verify the analysis results of system coverage ratio, we provide numerical examples in this section. The simulation parameters are shown in Table 1. Another important parameter of COM-NOMA is the coverage ratio of the first stage C_1, $C_1 \in (0.5, 0.8]$. Because lower C_1 than 0.5 means that each USU can not own one relay on average. There will not be enough SUs being selected as RUs of USUs in the second stage, especially when M_i is small. However, when C_1 becomes larger and close to 1, in the first stage, H_{2,s_2} becomes smaller. As the multicast rate is limited by the user with the worst channel condition. It is the more unfair to the user with good channel condition. To simplify simulation, it is assumed that two MGs have the same number of users, $M_1 = M_2$. In this paper, when the number of users of each MG is smaller than 60, it is considered as the small number of users. BS broadcasts the SINR threshold in each time slot. However, when it is larger than 100, it is considered as the large number of users. The SINR threshold is calculated according to (24) and BS broadcasts it at the beginning of the data transmission.

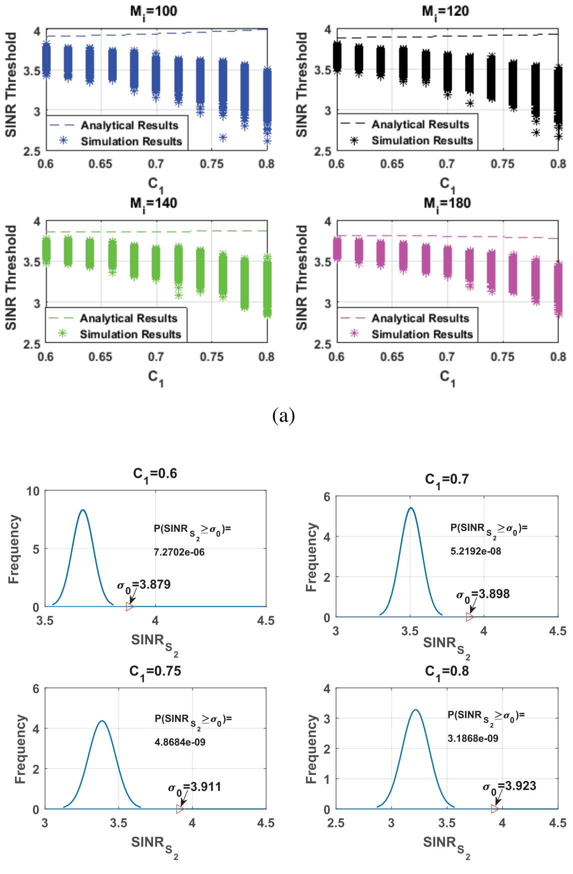
TABLE 1. Simulation parameters.

Parameter Name	Value
Coverage Radius of Cellular	$1000m$
Efficient Transmission Range	$100m$
System Bandwidth	$10MHz$
Transmission Power of BS	$34dBm$
Transmission Power of User Terminal	$17dBm$
Noise Power Spectrum Density	$-174dBm/Hz$
Path Loss Coefficient	4
Power Allocation Factor α_1	0.2
Power Allocation Factor α_2	0.8

A. PERFORMACE EVALUATION FOR SINR THRESHOLD

In Fig. 4, the analytical SINR threshold, which is computed through substituting (24) into $\sigma_0 = 2^{R_{\sigma_0}(C_1)} - 1$, and the first-stage multicast SINRs of 1000 time slots are shown, when the number of MG users is large. From Fig. 4(a), in different M_i, the analytical SINR threshold is not less than the maximum value of the multicast SINR in the first stage. It indicates that the analytical SINR threshold can guarantee the successful reception of the USUs who are in D2MD group. Fig. 4(b) shows the distribution curve of $SINR_{\mathcal{S}_2}$ in different C_i, when $M_i = 120$. The ordinate indicates the frequency of every $SINR_{\mathcal{S}_2}$ value. As can be seen from the figure, the distribution interval of $SINR_{\mathcal{S}_2}$ increases with the increase of C_1, which can also be verified by Fig. 4(a). The probabilities that $SINR_{\mathcal{S}_2}$ is greater than the SINR threshold δ_0 are given. They are all less than 10^{-5}, and decrease as C_1 increases. Thus, the analytical SINR threshold can reduce the frequency that BS computes and broadcasts threshold to the MG users in the instantaneous channel conditions. So the analytical SINR threshold is reasonable for USU to select RU when M_i is large.

$$\begin{aligned}P^A_{m_i} &= \frac{1}{2\pi}\int_{R_0}^{R} P_{m_i}(\gamma_u)\frac{2\gamma_u}{R^2}d\gamma_u\\ &\approx \frac{1}{2\pi}\int_{R_0}^{R} \arcsin\left(\frac{R_0}{\gamma_u}\right)\left(\int_{\gamma_u-R_0}^{\gamma_u+R_0}\frac{2\gamma_{m_i,u}}{R^2}d\gamma_{m_i,u}\right)\frac{2\gamma_u}{R^2}d\gamma_u\\ &= \frac{1}{2\pi}\int_{R_0}^{R} \arcsin\left(\frac{R_0}{\gamma_u}\right)\frac{4\gamma_u R_0}{R^2}\frac{2\gamma_u}{R^2}d\gamma_u\\ &= \frac{2R_0}{3\pi R^4}\left(R_0 R\sqrt{R^2-{R_0}^2} + 2R^3\arcsin\left(\frac{R_0}{R}\right) + {R_0}^3\ln\left[\left(1+\sqrt{1-\frac{{R_0}^2}{R^2}}\right)R\right] - \pi {R_0}^3 - {R_0}^3\ln R_0\right).\end{aligned} \tag{36}$$



(a)

(b)

FIGURE 4. SINR threshold in large number of MG user (a):SINR threshold versus C_1, in different number of users. (b):Probabilistic graphical of SINR threshold with $M_i = 120$, in different C_1.

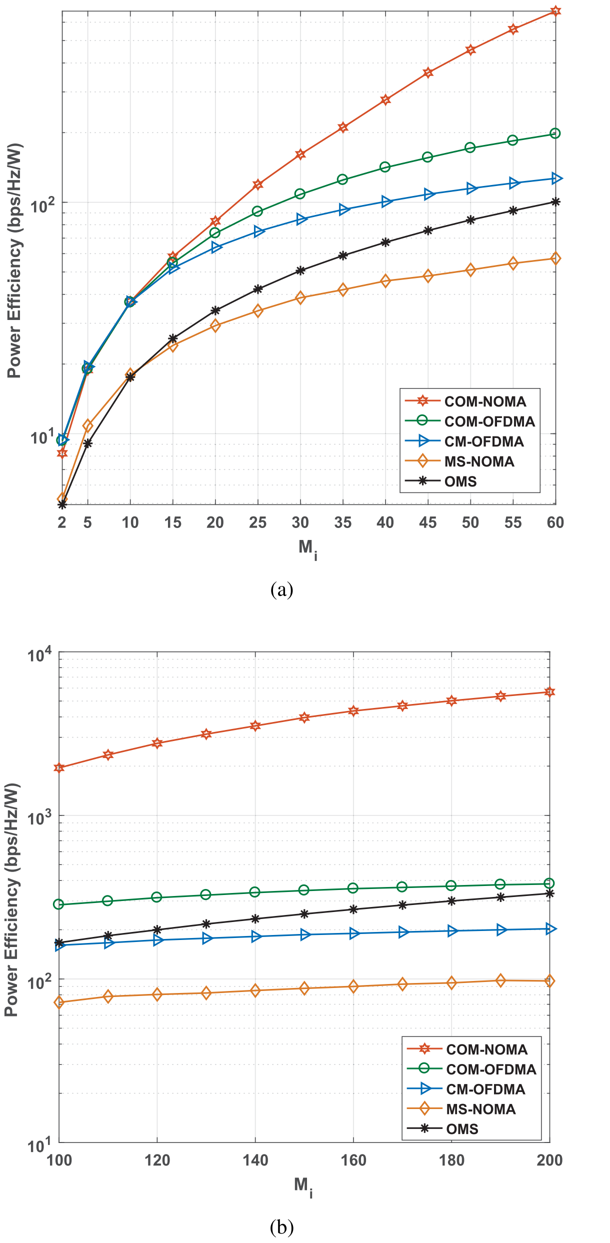
B. COMPARISON WITH EXISTING MULTICAST SCHEMES

Considering that NOMA is employed, we compare the proposed COM-NOMA scheme with four other multicast schemes.

COM-OFDMA[4] uses the same cooperation scheme with COM-NOMA scheme, but it is OFDMA-based, i.e., the two signals, x_1 and x_2, are separately transmitted on two orthogonal frequency bands. CM-OFDMA is a CM scheme proposed in [4]. The above three schemes support a fixed proportion of users in the first stage with coverage ratio C_1. But in CM-OFDMA scheme, USU combines receive signals from all SUs. When the rate of combined signals is higher than that of the first stage, the demand signal can be decoded. In the two OFDMA-based CM schemes, each MG collaborates in the group and the signals of two MGs are respectively transmitted on the two orthogonal frequency bands. Over the entire time slot, opportunistic multicast scheme (OMS) in [23] only supports a definite proportion of users who own good channel conditions, and its system coverage ratio is C_1. MS-NOMA is an NOMA-based multicast scheme serving all MG users, which is proposed in [16].

1) POWER EFFICIENCY COMPARISON

Due to different multicast schemes employ different number of RUs to forward data in the second stage, their total power consumption is different. For fairness, power efficiency of 5 kinds of multicast schemes are plotted in Fig. 5. The total power is denoted by $P_{tot} = \frac{T_1}{T}P + \frac{T_2}{T}P'$ [24], where P' is the total power of RUs in the second stage. From Fig. 5, it can be observed that COM-NOMA performs better than



(a)

(b)

FIGURE 5. Comparison of 5 multicast schemes in power efficiency, as a function of user number with $C_1 = 0.6, \alpha_2 = 0.8$. (a):$\sigma_0 = SINR_{S_2}$ in small number of MG users. (b):$\sigma_0 = 2^{R_{\sigma_0}(C_1)} - 1$ in large number of MG users.

[4] COM-OFDMA scheme is also proposed in this paper, in order to compare with COM-NOMA.

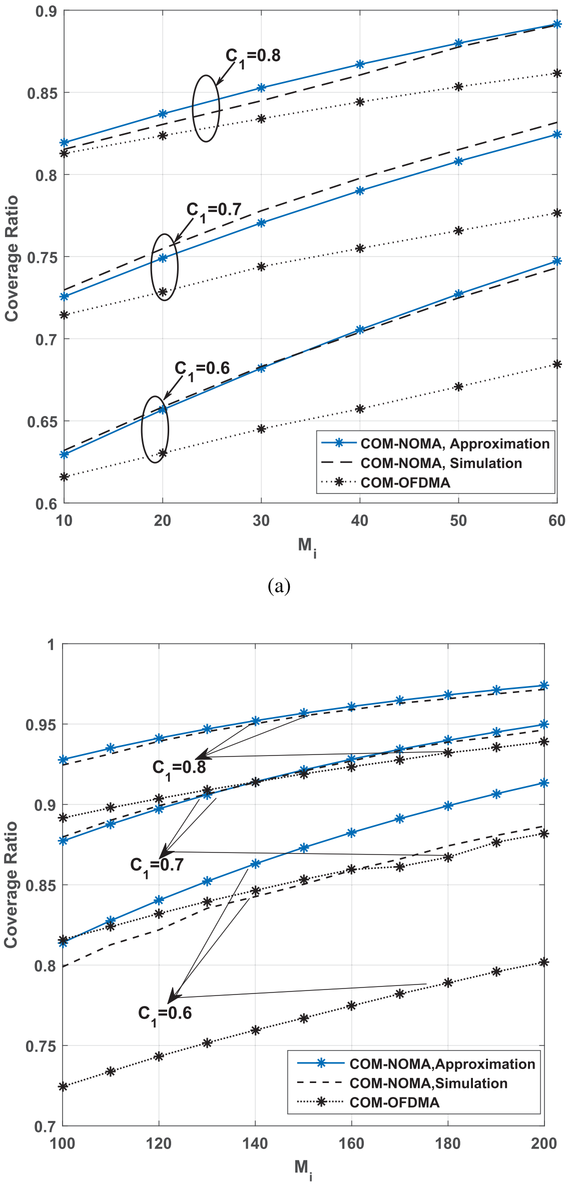
other multicast schemes when M_i is larger than 10. It is because that receiving both demand signals increases the probability that SUs forward signals to USUs, whichever MG they belong to. The cooperation between two MGs is achieved. When MG users are uniformly distributed in the cell, the number of candidate RUs is twice as many as other CM schemes. USU has more opportunities to be relayed by users. And each RU ensures that the USUs he (she) serves receive successfully. More D2MD groups are constituted in the second stage. Because the number of USUs is also larger than those two OFDMA-based CM schemes, the distance between USU and SU is relatively smaller. The number of members in each D2MD group is also larger. Multicast capacity is influenced by the number of users in each MG. Because $P_D << P$, the power consumption of BS is the main part of the total power consumption. In these three CM schemes, the power consumption of BS is the same. COM-NOMA achieves higher power efficiency than them.

In COM-OFDMA, due to the signals of two MGs are transmitted on the orthogonal frequency bands, different MGs can not cooperate with each other. The distance between users in the same MG may be larger than the radius of ETR. Therefore, intra-group cooperation also can not be performed well. BS only transmits data in the first half of the time slot, so its power efficiency is about twice that of OMS, when M_i is smaller 60. Although CM-OFDMA makes all SUs as relays to serve USUs in the second phase, severe path loss which is caused by the large distance between SU and USU impairs its performance. Therefore, their performance is lower than COM-NOMA scheme.

However, in Fig. 5(a), which simulates the power efficiency of COM-NOMA scheme with SINR threshold $\sigma_0 = SINR_{S_2}$ under small number of MG users. When M_i is smaller than about 10, the power efficiency of COM-NOMA is worse than COM-OFDMA. Because user density is lower in the cellular, SU hardly locates within the ETR of USU. Inter-group cooperation cannot be performed well. In order to guarantee efficient transmission of selected RUs, COM-NOMA scheme ignores the signal that makes reception SINR of USU smaller than threshold. This leads to less relays than CM-OFDMA to complete the transmission in the second stage. At the same time, $P_D << P$, the power of BS is the main factor of total power consumption, as M_i is small. The signal of one MG is the other MG's interference, so the performance of COM-NOMA is worse in comparison with CM-OFDMA. That also produces the poor performance of MS-NOMA. In Fig. 5(b), the threshold is $\sigma_0 = 2^{R_{\sigma_0}(C_1)} - 1$ for large M_i. CM-OFDMA has high power consumption due to all SUs are selected as relays. Its power efficiency is worse than that of OMS and COM-OFDMA.

2) COVERAGE RATIO COMPARISON

The comparison of coverage performance between COM-NOMA and COM-OFDMA under different C_1 is shown in Fig. 6. Since CM-OFDMA makes all SUs as the relays of USUs to forward data in the second stage regardless of device's ETR. For the sake of fairness, only the COM-OFDMA that utilizes the same relay selection scheme as the COM-NOMA is shown in the figure. Fig. 6(a) and Fig. 6(b) show the coverage performance under different number of MG users. COM-NOMA's coverage performance is superior to COM-OFDMA. Especially when $M_i = 200$, it increases by about 10%. This is because COM-NOMA implements inter-group cooperation between two MGs. When MG users are uniformly distributed in the cell, users of both MGs in the ETR of the USU can receive their flag messages. The probability that SUs are selected as relays increases,



(a)

(b)

FIGURE 6. System coverage ratio versus the number of MG users, in different C_1. (a): The number of MG users is small. SINR threshold $\sigma_0 = SINR_{S_2}$. (b): The number of MG users is large. SINR threshold $\sigma_0 = 2^{R_{\sigma_0}(C_1)} - 1$.

and the proposed relay selection scheme ensures that each USU covered by RU can successfully receive, which also increases the coverage ratio of system. It can be seen that the approximate and simulated coverage performance match, except for $C_1 = 0.6$ in Fig. 6(b). When $C_1 = 0.6$ in Fig. 6(b), approximate coverage ratio is larger than the simulated one. The difference between the approximate and simulated coverage ratio is mainly caused by the transmission in the second stage. It is because that in the calculation of the second-stage coverage ratio, the circular area actually covered by RU is approximated to its circumscribed sector field, which increases the coverage, i.e., the approximate P_{m_i} increases. The approximate expression of coverage ratio (35) is an increasing function of $P^A_{m_i}$ and $(M_1 + M_2)$. And when $C_1 = 0.6$, factor $(1 - C_1)$ is larger than $C_1 = 0.7$ and $C_1 = 0.8$. Therefore, the difference between approximate coverage ratio and the simulated one is obvious, when the number of MG users is large and $C_1 = 0.6$. But the error is still less than 3% of the simulated coverage ratio.

3) COMPLEXITY COMPARISON

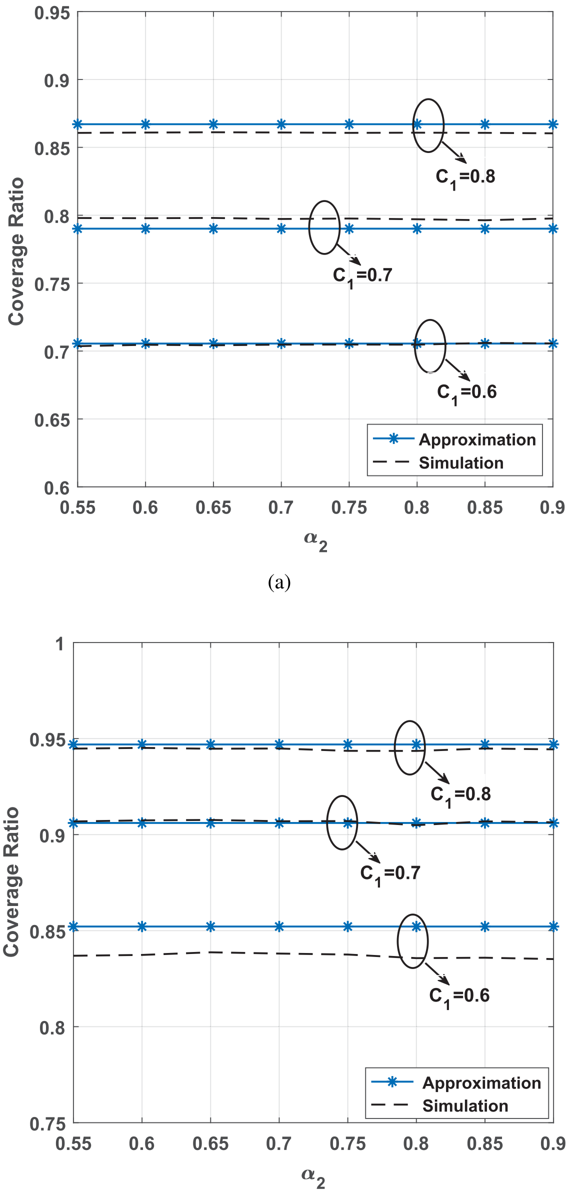
The complexity of the other four multicast schemes, i.e., COM-OFDMA, CM-NOMA, MS-NOMA and OMS, is also mainly influenced by the three steps in the Sec IV-B. The complexity analysis and comparison is shown in Tab. 2. From Tab. 2, the complexity of COM-NOMA is higher than that of the other schemes. It is because of that it performs RU selection for USU in the second stage after the first-stage opportunistic multicast with a fixed coverage ratio. But it improves the main drawback of OMS, namely, poor fairness. In OMS, USUs have no chance to recept successfully, due to their poor channel conditions, even though they have already paid for the content.

The complexity of COM-NOMA is determined by C_1, and the total number of users in the two MGs, $M_1 + M_2$. The complexity of those two OFDMA-based CM schemes, COM-OFDMA and CM-OFDMA, is determined by C_1 and M_i. It is because that they only perform collaboration within the single MG, unlike COM-NOMA, which also supports the collaboration between two MGs. As a result, in the third step, the amount of candidate RUs is less than the COM-NOMA for each MG.

Even though the complexity of the COM-NOMA is higher than other schemes, its energy efficiency is greatly improved. For example, in Figure 5(a), when $M_i = 60$, the complexity of COM-NOMA is twice that of COM-OFDMA and CM-OFDMA, but the power efficiency of COM-NOMA is 3 times that of COM-OFDMA, and 5 times that of CM-OFDMA. The complexity of COM-NOMA is 5 times that of MS-NOMA, but its power efficiency is 12 times that of MS-NOMA.

C. IMPACT OF PAF ON COVERAGE RATIO

As an important parameter of NOMA, the impact of different PAFs on the coverage ratio of COM-NOMA is simulated in Fig.7. Coverage ratio is nearly invariable under different PAFs. It is because that the coverage ratio C_1 is fixed (e.g., 0.6, 0.7, 0.8) in the first stage. The improvement of system coverage ratio is produced by the cooperation among MG users in the second stage. However, in the second stage, D2MD members (i.e., those USUs who are served by RUs) are distributed in the small ETRs of RUs. The small path loss caused by the samll distance between RU and D2MD member is the key factor to determine that the reception SINR of signal x_2 for D2MD members is larger than σ_0, even though the transmission power of RU is low. This also guarantees that each selected RU can make the USUs he (she) serves receive successfully. Therefore, coverage ratio of the second stage C_2



(a)

(b)

FIGURE 7. **System coverage ratio versus PAF. (a): The number of MG users is small and $M_i = 40$. SINR threshold $\sigma_0 = SINR_{S_2}$. (b): The number of MG users is large and $M_i = 120$. SINR threshold $\sigma_0 = 2^{R_{\sigma_0}(C_1)} - 1$.**

TABLE 2. Complexity analysis and comparison.

	Complexity of Step 1	Complexity of Step 2	Complexity of Step 3	Sum Complexity of Scheme
COM-NOMA	$\mathcal{O}(M_1\log_2 M_1)+\mathcal{O}(M_2\log_2 M_2)$	$\mathcal{O}(C_1M_1)$	$\mathcal{O}(C_1(1-C_1)(M_1+M_2)^2)$	$\mathcal{O}(C_1(1-C_1)(M_1+M_2)^2)$
COM-OFDMA	$\mathcal{O}(M_1\log_2 M_1)+\mathcal{O}(M_2\log_2 M_2)$	0	$\mathcal{O}(C_1(1-C_1)(M_1^2+M_2^2))$	$\mathcal{O}(C_1(1-C_1)(M_1^2+M_2^2))$
CM-OFDMA	$\mathcal{O}(M_1\log_2 M_1)+\mathcal{O}(M_2\log_2 M_2)$	0	$\mathcal{O}(C_1(1-C_1)(M_1^2+M_2^2))$	$\mathcal{O}(C_1(1-C_1)(M_1^2+M_2^2))$
MS-NOMA	$\mathcal{O}(M_1\log_2 M_1)+\mathcal{O}(M_2\log_2 M_2)$	M_1	0	$\mathcal{O}(M_1\log_2 M_1)+\mathcal{O}(M_2\log_2 M_2)$
OMS	$\mathcal{O}(M_1\log_2 M_1)+\mathcal{O}(M_2\log_2 M_2)$	0	0	$\mathcal{O}(M_1\log_2 M_1)+\mathcal{O}(M_2\log_2 M_2)$

does not change significantly with α_2, which also validates the proof in (35) and (36) that coverage ratio is independent of PAF. Therefore, the coverage ratio of COM-NOMA does not change significantly with α_2.

VI. CONCLUSIONS AND FUTURE WORK

In this paper, an NOMA-based CM scheme is proposed to achieve inter-group cooperation between two MGs, which is termed as COM-NOMA. The signals of two MGs are transmitted on the same licensed spectrum simultaneously in NOMA way. Thus, every SU gets the both signals. SUs in one MG have the chance to be selected as the RUs to serve the USUs in the other MG. To guarantee that the USUs served by RUs can successfully decode, SINR threshold for RU selection is given based on the extreme value theory and the statistical property of the first-stage multicast SINR, when the number of MG users is large (e.g., 100). It is different from the case that the number of MG users is small (e.g., 60). BS doesn't have to calculate and broadcast the first-stage multicast SINR in every time slot. We have further investigated the coverage ratio of COM-NOMA. The approximated close-form expression of system coverage ratio is derived. According to the analytical result, it is mainly impacted by number the number of multicast users. The PAF, which is an important parameter of NOMA, scarcely contributes to coverage ratio. Simulation and numerical computations have been carried out to verify the performance of COM-NOMA. It shows that power efficiency enhances when the number of users is larger than 10. Moreover, the coverage ratio enhances in comparison with COM-OFDMA, especially when $M_i = 200$. It increases by about 10%. And the analytical coverage ratio and the simulated one match well.

In the future, the boundary of the total number of MG users will be investigated to choose proper SINR threshold, which is not investigated in this paper. In addition, the proposed COM-NOMA schemes is not energy-efficient, because one USU can select multiple SUs as his (her) RUs (especially in the area close to BS) and every RU can make sure successful reception. According to the user density in the cell and the distance between user and BS, the optimal RU is selected to save energy.

REFERENCES

[1] G. Araniti, P. Scopelliti, G. M. Muntean, and A. Iera, "A hybrid unicast-multicast network selection for video deliveries in dense heterogeneous network environments," *IEEE Trans. Broadcast.*, to be published.

[2] S. Pizzi, M. Condoluci, G. Araniti, A. Molinaro, A. Iera, and G.-M. Muntean, "A unified approach for efficient delivery of unicast and multicast wireless video services," *IEEE Trans. Wireless Commun.*, vol. 15, no. 12, pp. 8063–8076, Dec. 2016.

[3] S. Carson et al., "Ericsson mobility report: On the pulse of the networked society," Ericsson, Stockholm, Sweden, Jun. 2015. [Online]. Available: https://www.ericsson.com/assets/local/mobility-report/documents/2015/ericsson-mobility-report-june-2015.pdf

[4] F. Hou, L. X. Cai, P. H. Ho, X. Shen, and J. Zhang, "A cooperative multicast scheduling scheme for multimedia services in IEEE 802.16 networks," *IEEE Trans. Wireless Commun.*, vol. 8, no. 3, pp. 1508–1519, Mar. 2009.

[5] J. Lee, Y. M. Lim, K. Kim, S. G. Choi, and J. K. Choi, "Energy efficient cooperative multicast scheme based on selective relay," *IEEE Commun. Lett.*, vol. 16, no. 3, pp. 386–388, Mar. 2012.

[6] H. V. Zhao and W. Su, "Cooperative wireless multicast: Performance analysis and power/location optimization," *IEEE Trans. Wireless Commun.*, vol. 9, no. 6, pp. 2088–2100, Jun. 2010.

[7] O. Alay, T. Korakis, Y. Wang, and S. Panwar, "Layered wireless video multicast using directional relays," in *Proc. IEEE ICASSP*, Mar. 2008, pp. 2149–2152.

[8] Y. Zhou, H. Liu, Z. Pan, L. Tian, J. Shi, and G. Yang, "Two-stage cooperative multicast transmission with optimized power consumption and guaranteed coverage," *IEEE J. Sel. Areas Commun.*, vol. 32, no. 2, pp. 274–284, Feb. 2014.

[9] 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.

[10] H. Zhang, Y. Qiu, K. Long, G. K. Karagiannidis, X. Wang, and A. Nallanathan, "Resource allocation in NOMA-based fog radio access networks," *IEEE Wireless Commun.*, vol. 25, no. 3, pp. 110–115, Jun. 2018.

[11] F. Fang, H. Zhang, J. Cheng, S. Roy, and V. C. M. Leung, "Joint user scheduling and power allocation optimization for energy-efficient NOMA systems with imperfect CSI," *IEEE J. Sel. Areas Commun.*, vol. 35, no. 12, pp. 2874–2885, Dec. 2017.

[12] L. Lv, J. Chen, Q. Ni, and Z. Ding, "Design of cooperative non-orthogonal multicast cognitive multiple access for 5G systems: User scheduling and performance analysis," *IEEE Trans. Commun.*, vol. 65, no. 6, pp. 2641–2656, Jun. 2017.

[13] L. Yang, J. Chen, Q. Ni, J. Shi, and X. Xue, "NOMA-enabled cooperative unicast-multicast: Design and outage analysis," *IEEE Trans. Wireless Commun.*, vol. 16, no. 12, pp. 7870–7889, Dec. 2017.

[14] Y. Chen, L. Wang, and B. Jiao, "Cooperative multicast non-orthogonal multiple access in cognitive radio," in *Proc. IEEE ICC*, May 2017, pp. 1–6.

[15] Z. Ding, Z. Zhao, M. Peng, and H. V. Poor, "On the spectral efficiency and security enhancements of NOMA assisted multicast-unicast streaming," *IEEE Trans. Commun.*, vol. 65, no. 7, pp. 3151–3163, Jul. 2017.

[16] Y. Zhang, X. Wang, D. Wang, Q. Zhao, and Y. Zhang, "Joint transmission scheme for two multicast groups based on NOMA," in *Proc. IEEE PIMRC*, Oct. 2017, pp. 1–6.

[17] O. Alay, P. Liu, Y. Wang, E. Erkip, and S. S. Panwar, "Cooperative layered video multicast using randomized distributed space time codes," *IEEE Trans. Multimedia*, vol. 13, no. 5, pp. 1127–1140, Oct. 2011.

[18] Z. Mo, W. Su, S. Batalama, and J. D. Matyjas, "Analysis and optimization of distributed cooperative multicast for wireless multimedia networks," in *Proc. IEEE ICC*, Jun. 2014, pp. 5573–5579.

[19] B. Niu, H. Jiang, and H. V. Zhao, "A cooperative multicast strategy in wireless networks," *IEEE Trans. Veh. Technol.*, vol. 59, no. 6, pp. 3136–3143, Jul. 2010.

[20] L. Feng *et al.*, "Resource allocation for 5G D2D multicast content sharing in social-aware cellular networks," *IEEE Commun. Mag.*, vol. 56, no. 3, pp. 112–118, Mar. 2018.

[21] W. Cheng, X. Zhang, and H. Zhang, "Optimal power allocation with statistical QoS provisioning for D2D and cellular communications over underlaying wireless networks," *IEEE J. Sel. Areas Commun.*, vol. 34, no. 1, pp. 151–162, Jan. 2016.

[22] N. Lee, X. Lin, J. G. Andrews, and R. W. Heath, Jr., "Power control for D2D underlaid cellular networks: Modeling, algorithms, and analysis," *IEEE J. Sel. Areas Commun.*, vol. 33, no. 1, pp. 1–13, Jan. 2015.

[23] T.-P. Low, M.-O. Pun, Y.-W. P. Hong, and C.-C. J. Kuo, "Optimized opportunistic multicast scheduling (OMS) over heterogeneous cellular networks," *IEEE Trans. Wireless Commun.*, vol. 9, no. 2, pp. 791–801, Feb. 2010.

[24] Y. Zhou, H. Liu, Z. Pan, L. Tian, and J. Shi, "Cooperative multicast with location aware distributed mobile relay selection: Performance analysis and optimized design," *IEEE Trans. Veh. Technol.*, vol. 66, no. 9, pp. 8291–8302, Sep. 2017.

YUFANG ZHANG received the B.S. degree from the Changchun University of Science and Technology, China, in 2013. She is pursuing the Ph.D. degree with the Key Laboratory of Universal Wireless Communications, Ministry of Education, Beijing University of Posts and Telecommunications, Beijing, China. Her research interests include non-orthogonal multiple access, wireless multicast, and cooperative communication.

XIAOXIANG WANG received the B.S. degree in physics from Qufu Normal University, Qufu, China, in 1991, the M.S. degree in information engineering from East China Normal University, Shanghai, China, in 1994, and the Ph.D. degree in electronic engineering from the Beijing Institute of Technology, Beijing, China, in 1998. In 1998, she joined the School of Information and Communication Engineering, Beijing University of Posts and Telecommunications. From 2001 to 2002, she was a Visiting Fellow with the School of Electrical Engineering and Information Technology, Vienna University of Technology, Vienna, Austria. From 2010 to 2011, she was a Visiting Fellow with the Department of Electrical and Computer Engineering, North Carolina State University, Raleigh. Her research interests include communications theory and signal processing, with specific interests in cooperative communications, multiple-input–multiple-output systems, multimedia broadcast/multicast service systems, and resource allocation.

DONGYU WANG received the B.S. and M.S. degrees from Tianjin Polytechnic University, China, in 2008 and 2011, respectively, and the Ph.D. degree from the Beijing University of Posts and Telecommunications, China, in 2014. Since 2014, he has been a Post Scholar with the Department of Biomedical Engineering, Chinese PLA General Hospital, Beijing. His research interests include device to device communication, multimedia broadcast/multicast service systems, resource allocation, theory and signal processing, with specific interests in cooperative communications, and MIMO systems.

YIBO ZHANG received the B.S. degree from Henan University, Kaifeng, China, in 2013, and the M.S. degree from Henan Normal University, Xinxiang, China, in 2015. She is currently pursuing the Ph.D. degree with the Key Laboratory of Universal Wireless Communications, Ministry of Education, Beijing University of Posts and Telecommunications, Beijing, China. His research interests include non-orthogonal multiple access, wireless multicast, and cooperative communication.

QIANG ZHAO received the B.S. and M.S. degrees from the Aeronautical University of Air Force, Changchun, China, in 2011 and 2013, respectively, and the Ph.D. degree from Space Engineering University, Beijing, China, in 2018. His research interests include radio interferometry and satellite signal processing.

QIAN DENG received the B.S. and M.S. degrees in communication engineering from the Guilin University of Electronic Technology, China, in 2006 and 2009, respectively. She is currently pursuing the Ph.D. degree with the Beijing University of Posts and Telecommunications, China. Her current research interests include low complexity linear precoding in massive MIMO system and distributed beamforming in cognitive wireless network.