Massive MIMO Cognitive Cooperative Relaying

Son Dinh[1], Hang Liu[1], Feng Ouyang[2]

[1] The Catholic University of America, Washington, DC, USA
[2] Johns Hopkins Applied Physics Laboratory, Laurel, MD, USA

Abstract. This paper proposes a novel cognitive cooperative transmission scheme by exploiting massive multiple-input multiple-output (MMIMO) and non-orthogonal multiple access (NOMA) radio technologies, which enables a macrocell network and multiple cognitive small cells to cooperate in dynamic spectrum sharing. The macrocell network is assumed to own the spectrum band and be the primary network (PN), and the small cells act as the secondary networks (SNs). The secondary access points (SAPs) of the small cells can cooperatively relay the traffic for the primary users (PUs) in the macrocell network, while concurrently accessing the PUs' spectrum to transmit their own data opportunistically through MMIMO and NOMA. Such cooperation creates a "win-win" situation: the throughput of PUs will be significantly increased with the help of SAP relays, and the SAPs are able to use the PUs' spectrum to serve their secondary users (SUs). The interplay of these advanced radio techniques is analyzed in a systematic manner, and a framework is proposed for the joint optimization of cooperative relay selection, NOMA and MMIMO transmit power allocation, and transmission scheduling. Further, to model network-wide cooperation and competition, a two-sided matching algorithm is designed to find the stable partnership between multiple SAPs and PUs. The evaluation results demonstrate that the proposed scheme achieves significant performance gains for both primary and secondary users, compared to the baselines.

Keywords: Massive MIMO, Non-orthogonal multiple access, Dynamic spectrum access, Relay selection, Cooperative spectrum sharing.

1 Introduction

Mobile traffic is growing at a very rapid rate. Massive multiple-input, multiple-output (MMIMO) and non-orthogonal multiple access (NOMA) are two essential enabling technologies for next-generation (5G & beyond) mobile networks to achieve necessary performance improvement in spectrum efficiency and network capacity for meeting ever increasing user demands. Traditional MIMO networks typically use a few of antennas to transmit and receive signals. Massive MIMO (MMIMO), on the other hand, is a MIMO system using an antenna array with a large number of elements at the base stations (BSs) or access points (APs) [1, 2]. Advanced signal processing techniques can be employed to leverage the large number of antennas and concurrently generate multiple directional signal beams, each focusing a great amount of signal energy on an

intended mobile user (MU). Beamforming enables the BS/AP to transmit/receive multiple signal beams simultaneously to/from multiple MUs on the same frequency channel with a high signal gain. The more antenna elements the BS/AP is equipped with, the more possible signal paths and the higher total throughput. The emerging 3GPP 5G New Radio (NR) standards [3, 4] support MMIMO in both mmWave bands and sub 6GHz bands with up to 64 logical antenna ports, and the number of antenna elements is expected to increase in the future standard releases. NOMA is another technique to improve spectrum efficiency and network throughput [5, 6]. With NOMA, a user receiving the superposition transmission with its own signal sent in lower power can decode the stronger signal components for other users and then cancel them out to get its own signal, thus yields a significant spectral efficiency gain over conventional orthogonal multiple access. These new radio techniques can potentially significantly enhance the network performance and distinguish 5G systems from 4G systems.

In addition, 5G NR will support services with different spectrum licensing terms [7], including exclusive-use licensed spectrum, shared spectrum, and unlicensed spectrum. In particular, dynamic access to shared spectrum through cognitive radio (CR) capability can make more efficient use of spectrum, alleviate the spectrum scarcity problem, and provides new services. For example, non-operator organizations can use shared spectrum to deploy private networks in public venues, workplaces, or industrial facilities, which will unlock opportunities for innovative deployment models and take advantage of 5G technology to extend mobile networking ecosystem.

It is vital to have efficient and reliable mechanisms to optimize dynamic spectrum access (DSA) to the shared spectrum. There can be different models for DSA [8]. In interweave or underlay DSA models that most existing research focused on, unlicensed secondary users (SUs) of spectrum can access the licensed spectrum bands of primary users (PUs) to transmit data only when the PUs are not using the spectrum or when the interference from the SUs are tolerable by the PUs, i.e. below certain threshold, through techniques such as spectrum sensing and interference management. Alternatively, the PUs and SUs can cooperate in DSA, termed cooperative DSA [9, 10], also known as cooperative cognitive radio network (CCRN) model, to achieve flexible spectrum sharing. The cooperative spectrum sharing can perform more efficiently than uncooperative shared access and benefit both parties. Novel network architecture and protocols are needed to facilitate the cooperation. Specifically, it is worth to investigate whether joint optimization of various elements in the network system is possible and design efficient algorithms to leverage advanced physical-layer technologies such as MMIMO and NOMA in cooperative cognitive radio networks for significant performance gains and new network functionalities.

In this paper, we propose a novel cooperative transmission scheme of PUs and SUs by exploring new radio technologies such as MMIMO and NOMA in dynamic spectrum access and sharing. We study a deployment scenario consisting of a cellular macrocell network and multiple cognitive small cells, in which the incumbent macrocell network owns the spectrum band and is the primary network (PN), and the small cells act as the secondary networks (SNs). The macrocell network serves a group of primary users, and small cell networks serve their own secondary users. The secondary access points (SAPs) of the small cells equip with cognitive radio capability with MMIMO beamforming and NOMA technologies. A SAP can dynamically access the spectrum owned by the macrocell network to help the incumbent BS to relay the primary traffic

to the PUs while simultaneously transmit its own data with MMIMO and NOMA. Such cooperation creates a "win-win" situation: the throughput of PUs will be significantly increased with the assistance of SAP relays, and the SAPs can serve their SUs opportunistically. In this way, the dynamic spectrum access by the small cells will not congest the licensed spectrum, but improve the performance of the incumbent primary network. The interplay of these advanced radio techniques is analyzed in a systematic manner, and a framework for the joint optimization of relay selection, NOMA and MMIMO transmit power allocation, and transmission scheduling is proposed and investigated. Further, to model network-wide cooperation and competition, a two-sided matching algorithm is designed to find stable partnership between the SAPs and PUs in the network. The evaluation results demonstrate that the proposed scheme greatly improves the utilities of both primary and secondary users.

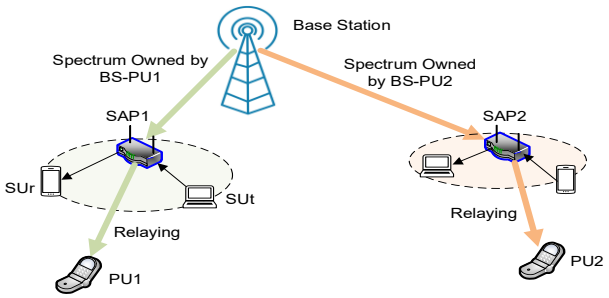


Fig. 1. A scenario for cognitive cooperative relaying with MMIMO and NOMA.

2 System Model

As shown in Fig. 1, there exist a group of small cells in the coverage area of an incumbent macrocell base station (BS) that is the owner of a spectrum band. The incumbent BS serves a number of PUs. We assume that a PU is allocated a licensed sub-channel for data delivery in a time slot via orthogonal frequency-division multiplexing (OFDM). Thus, we define a link between the macrocell BS and PU as the primary link (PL). For simplicity, we assume the incumbent BS and PUs are equipped with a single antenna and no NOMA capability.

Each small cell SAP is assumed to have MMIMO and NOMA interference cancellation capability, which can dynamically access to the sub-channels in the licensed spectrum of the incumbent PUs to serve its SUs opportunistically. Under the proposed cooperative DSA framework, a SAP can dynamically relay the traffic for the PU, while borrowing PU's sub-channel to transmit/receive the secondary data to/from its SUs using its MIMO and NOMA capabilities. We design the algorithms for a PU to select a SAP as relay and the SAP to control the power for transmitting SU data in the small cell and relaying PU data to optimize overall system performance. The SUs in a small cell are served opportunistically. For system fairness and simplicity, we allow one PU at most has one small cell SAP as its relay in a time slot. A SAP can only help one PU and access the PU channel that it is in cooperation with. In addition, we consider the downlink data communication from the macrocell BS to the PUs, whereas both uplink

and downlink transmissions are considered for the SN. For the uplink from PU to the macrocell BS, symmetric analysis can be applied.

If a PU_i does not have a relay during a transmission time slot t, the BS will directly send data to PU_i on the subchannel allocated to PU_i. If a SAP_j acts as a relay for a PU_i in a time slot t, we call SAP_j and PU_i forms a partnership. Let $\mathcal{S}_j$ denote a set of SUs in small cell j associated to SAP_j. Thus, the time slot t is divided into two subslots as shown in Fig. 2. In subslot 1, the BS will transmit PU_i's data on the subchannel allocated to PU_i, and the partner SAP_j will receive the PU_i's data. Meanwhile, SAP_j will schedule K SUs in its small cell, $SU_k, k \in \mathcal{S}_j$ to transmit secondary uplink traffic and utilize its MMIMO beamforming and NOMA signal cancellation capabilities to receive the secondary uplink traffic, while receiving the primary data. In subslot 2, the SAP_j forwards the primary data to PU_i and also sends K downlink secondary traffic beams to its SUs in the small cell, $SU_k, k \in \mathcal{S}_j$, with MMIMO and NOMA. We will consider the case where an SAP is equipped with M antennas ($M > K$) and a SU has a single antenna.

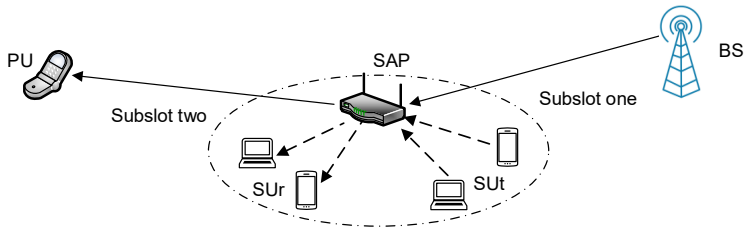


Fig. 2. MMIMO NOMA transmissions.

Assume the slot duration is T, subslot 1 duration δT $(0 \leq \delta \leq 1)$, and subslot 2 size $(1-\delta)T$. In subslot 1, an SAP_j accesses the subchannel allocated to its partner PU_i and receives the signal sent from the incumbent BS to PU_i, along with the uplink signals from K SUs. Let y_m be the baseband signal output at the m-th element of the SAP_i antenna array. The M x 1 signal vector at the array output, $\mathbf{y_j} = [y_{1j}, y_{2j}, \dots, y_{mj}, \dots y_{Mj}]^T$ can be represented by [11]

$$\mathbf{y_j} = \mathbf{h_{bj}} x_{bi} + \mathbf{H_j x_j} + \mathbf{n_j} \tag{1}$$

where x_{bi} is the primary message sent from the incumbent BS to PU_i and the elements of K x 1 vector $\mathbf{x_j} = [x_{1j}, x_{2j}, \dots, x_{kj}, \dots x_{Kj}]^T$, represent the messages transmitted from each of K SUs in the small cell $\mathcal{S}_j$. $\mathbf{n_j}$ is noise, and $\mathbf{H_j}$ is the $M \times K$ channel matrix between the SAP_i and its SUs. The column of the channel matrix, $\mathbf{H_j} = [\mathbf{h_{1j}}, \dots, \mathbf{h_{kj}}, \dots \mathbf{h_{Kj}}]$, represent the channels or spatial signatures associated with each SU. The channel vector with a linear antenna array can be modelled as [12]

$$\mathbf{h_{kj}} = \frac{\beta_{kj}}{1+b_{kj}^{e}} [1,\ e^{-j\pi\varphi_{kj}}, \dots,\ e^{-j\pi(M-1)\varphi_{kj}}] \tag{2}$$

where b_{kj} is the distance between the SAP_j and SU k, $k \in \mathcal{S}_j$, e is the pathloss exponent, φ_{kj} is the normalized direction, and β_{kj} is the fading attenuation coefficient. $\mathbf{h_{bj}}$ is the channel vector between the incumbent BS and partner SAP_i that can be modeled in the

same way. We are considering a flat Rayleigh block-fading channel model [10] to simplify our problem description, with which the channel is invariant and flat within each slot, but generally varying over the slots. For a large M value in MMIMO, a simple conjugate beamforming structure, i.e. a maximum-ratio-combining (MRC) beamformer with $\mathbf{w_{kj}} = \mathbf{h_{kj}^{H}}$ at SAP can yield good performance [13]. The symbol from the k-th user can thus be decoded by applying $\mathbf{w_{kj}} = \mathbf{h_{kj}^{H}}$ to the array output:

$$\hat{x}_{kj} = \mathbf{w_{kj}}\mathbf{y_j} = \mathbf{h_{kj}^{H}}\mathbf{h_{bj}}\mathrm{x_{bi}} + \mathbf{h_{kj}^{H}}\mathbf{H_j}\mathbf{x_j} + \mathbf{h_{kj}^{H}}\mathbf{n_j} \tag{3}$$

SAP decodes the message from each of the SUs by treating the signal from the incumbent BS and other SUs as interference, with the following signal-to-interference-plus-noise ratio (SINR):

$$SINR_{kj} = \frac{|\boldsymbol{h}_{kj}^{H}\boldsymbol{h}_{kj}|^2\alpha_{kj}^2}{\sum_{l\in\mathcal{S}_j\setminus k}|\boldsymbol{h}_{kj}^{H}\boldsymbol{h}_{lj}|^2\alpha_{lj}^2+|\boldsymbol{h}_{kj}^{H}\boldsymbol{h}_{bj}|^2\alpha_{bi}^2+\sigma_j^2} \tag{4}$$

where α_{lj}^2 is the signal transmit power of SU_l, $l \in \mathcal{S}_j$, and α_{bi}^2 is the power that the BS transmits the primary data to $\mathrm{PU_i}$, and σ_j^2 is the noise power. Then the achievable data rate from SU k to SAP j can then be expressed as $R_{kj} = B\log_2(1 + SINR_{kj})$. The data throughput from SU k to SAP j during subslot 1 can then be expressed as

$$C_{kj} = \delta TR_{kj} = \delta \mathrm{T}B\log_2(1 + SINR_{kj}) \tag{5}$$

After $\mathrm{SAP_j}$ decodes the messages from its SUs, it subtracts these messages from the superposed signal it received by carrying out successive interference cancellation (SIC) [5, 6], and decode the information from the incumbent BS for $\mathrm{PU_i}$ using a MRC beamformer with $\mathbf{w_{bj}} = \mathbf{h_{bj}^{H}}$. The SINR for decoding the $\mathrm{PU_i}$ information is given by

$$SINR_{bj} = \frac{|\boldsymbol{h}_{bj}^{H}\boldsymbol{h}_{bj}|^2\alpha_{bi}^2}{\sigma_j^2} \tag{6}$$

Note that if there is no NOMA SIC, the SINR for the $\mathrm{PU_i}$ message received by relay $\mathrm{SAP_j}$ is less due to the interference of the SUs, which is

$$SINR_{bj} = \frac{|\boldsymbol{h}_{bj}^{H}\boldsymbol{h}_{bj}|^2\boldsymbol{\alpha}_{bi}^2}{\sum_{l\in\boldsymbol{\mathcal{S}_j}}|\boldsymbol{h}_{bj}^{H}\boldsymbol{h}_{lj}|^2\boldsymbol{\alpha}_{lj}^2+\sigma_j^2} \tag{7}$$

By applying SIC, the interference to the PU information is removed before decoding so that its SINR is improved. The achievable data rate for the link from the incumbent BS to $\mathrm{SAP_j}$ can be expressed as: $R_{bj} = B\log_2(1 + SINR_{bj})$. The corresponding data throughput for the link from the incumbent BS to $\mathrm{SAP_j}$ in subslot 1 is

$$C_{bj} = \delta TR_{bj} = \delta TB\log_2(1 + SINR_{bj}) \tag{8}$$

It is possible to apply SIC in decoding the messages from the SUs. However, this will significantly increase the signal processing complexity and the SUs are served opportunistically, thus we only use SIC for PU message decoding to improve PU's performance. The SU message decoding depends on MMIMO beamforming. In addition,

we assume each SU uses a fixed power to transmit its uplink traffic as the uplink power control introduces a large overhead and complexity.

In subslot 2, the SAP_j relays the primary data x_{ji} to the PU_i and simultaneously transmits K secondary downlink messages, $\mathbf{x_{jd}} = [x_{j1}, x_{j2}, \dots, x_{jk}, \dots x_{jK}]^T$, one message to a SU, on the PU_i's subchannel with MMIMO beamforming and NOMA. Similar analysis can be performed. Let $\hat{x}_{ji}$ be the signal received by the PU_i and vector $\hat{\mathbf{x}}_{\boldsymbol{jd}} = [\hat{x}_{j1}, \hat{x}_{j2}, \dots \hat{x}_{jk}, \dots \hat{x}_{jK}]^T$ contain the signals received at each of the SUs, respectively, which can be described by

$$\hat{\mathrm{x}}_{ji} = \mathbf{h_{ji}^H w_{ji}} x_{ji} + \mathbf{h_{ji}^H W_j x_{jd}} + \mathrm{n_i} \tag{9}$$

$$\hat{\mathbf{x}}_{\boldsymbol{jd}} = \mathbf{H_j^H w_{ji}} x_{ji} + \mathbf{H_j^H W_j x_{jd}} + \mathbf{n_d} \tag{10}$$

where $\mathbf{w_{ji}}$ is the M x 1 MMIMO precoding vector applied to x_{ji} before transmitting it by the M antenna elements of the SAP_j to PU_i, and $\mathbf{W_j}$ is the $M \times K$ MMIMO precoding matrix applied to the secondary signal vector sent to its SUs by the SAP_j for the transmit beamforming. The downlink channel matrix $\mathbf{H_j^H}$ from the SAP_j to the SUs is considered as the conjugate transpose of the uplink channel matrix due to channel reciprocity. $\mathbf{h_{ji}}$ is the channel vector between the SAP_j and PU_i, and $\mathrm{n_i}$ is noise. We consider that the maximum ratio transmission (MRT) precoding [11] is used to transmit the primary and secondary messages to individual users, that is, $\mathbf{w_{ji}} = \mathbf{h_{ji}}$ and $\mathbf{W_j} = \mathbf{H_j}$.

An incumbent PU_i receives a superposition of the messages for itself as well as the SUs. It treats the SUs' information as noise and decodes its own message with the following SINR:

$$SINR_{ji} = \frac{\mathbf{h_{ji}^H h_{ji}} \alpha_{ji}^2}{\sum_{l \in \mathcal{S}_j} \mathbf{h_{ji}^H h_{lj}} \alpha_{jl}^2 + \sigma_i^2} \tag{11}$$

The achievable data rate of an incumbent PU_i during subslot 2 is then $R_{ji} = B \log_2(1 + SINR_{ji})$, and the corresponding PU_i throughput in subslot 2 is

$$C_{ji} = (1 - \delta) T R_{ji} = (1 - \delta) T B \log_2(1 + SINR_{ji}) \tag{12}$$

Assume that SUs have NOMA capability. After a SU receives the superposed signal, it may try two approaches to obtain its message, depending on its MMIMO channel state and SAP_j's power allocation strategy:

1) a SU_k, $k \in \mathcal{S}_j$, can try to decode the PU_i's message and then use SIC to subtract this message from its observation, and finally decode its own information. For the PU_i message decoding at SU_k, the SINR is given as:

$$SINR_{jk\_ji} = \frac{\mathbf{h_{kj}^H h_{ji}} \alpha_{ji}^2}{\sum_{l \in \mathcal{S}_j} \mathbf{h_{kj}^H h_{lj}} \alpha_{jl}^2 + \sigma_k^2} \tag{13}$$

The data rate for PU_i is R_{ji} and let $\varepsilon_{ji} = (2^{R_{ji}/B} - 1)$. If $SINR_{jk\_ji} \geq \varepsilon_{ji}$, SIC can be carried out successfully at SU k and the SINR for decoding its own message is given by

$$SINR_{jk} = \frac{\mathbf{h}_{\mathrm{kj}}^{\mathrm{H}}\mathbf{h}_{\mathrm{kj}}\alpha_{jk}^{2}}{\sum_{l\in\mathcal{S}_j\setminus k}\mathbf{h}_{\mathrm{kj}}^{\mathrm{H}}\mathbf{h}_{lj}\alpha_{jl}^{2}+\sigma_k^2} \tag{14}$$

2) If SU_k cannot successfully decode the PU_i signal, i.e. $SINR_{jk\_ji} < \varepsilon_{ji}$, it will decode its own message directly by treating PU_i's information as noise. The SINR of SU_k signal is then

$$SINR_{jk} = \frac{\mathbf{h}_{\mathrm{kj}}^{\mathrm{H}}\mathbf{h}_{\mathrm{kj}}\alpha_{jk}^{2}}{\mathbf{h}_{\mathrm{kj}}^{\mathrm{H}}\mathbf{h}_{\mathrm{ji}}\alpha_{ji}^{2}+\sum_{l\in\mathcal{S}\setminus k}\mathbf{h}_{\mathrm{kj}}^{\mathrm{H}}\mathbf{h}_{lj}\alpha_{jl}^{2}+\sigma_k^2} \tag{15}$$

The achievable data rate for SU_k is thus $R_{jk} = B\,log_2(1 + SINR_{jk})$, and the corresponding data throughput for SU_k in subslot 2 is

$$C_{jk} = (1-\delta)TR_{jk} = (1-\delta)TB\log_2(1+SINR_{jk}) \tag{16}$$

Moreover, If the incumbent BS transmits data to PU_i directly on its subchannel without cooperative relaying, the achievable rate is a function of the BS transmit power α_{bi}^2 and the complex channel gain h_{bi} between BS and PU_i, which can be expressed as [11], $R_{bi,dir} = Blog_2(1 + \frac{\alpha_{bi}^2}{\sigma_i^2}|h_{bi}|^2)$ where σ_i^2 is the noise power. The PU_i throughput without the cooperative relaying in time slot T is

$$C_{bi,dir} = TR_{bi,dir} = TBlog_2(1 + \frac{\alpha_{bi}^2}{\sigma_i^2}|h_{bi}|^2) \tag{17}$$

3 System Optimization

From the above analysis, we can see that a set of strategies affect the achievable throughput of PU and SAP transmissions, including (i) a PU should decide whether to use its frequency channel for direct transmission from BS to PU, or for SAP relaying. (ii) In the latter case, the best MMIMO-NOMA SAP relay for a PU should be selected. (iii) After a SAP relay is selected, how are the resources shared in the PU and SU data transmissions? That is, the MMIMO-NOMA relay transmission and power allocation strategies should be decided, including the size of subslots 1 and 2 as well as the SAP power allocation for transmitting PU data and SU data. We model the system of multiple PUs and multiple SAPs as a two-side matching problem, and study the relay selection, relay transmission, and power allocation strategies for overall system optimization.

3.1 System Utility Maximization

Let $\mathcal{P}$ denote the set of PU links and $\mathcal{S}$ the set of MMIMO-NOMA small cells each led by a SAP. We define the utility that each party can earn as its throughput, which is a function of relay selection, subslot partition, transmit power allocation, and MMIMO-NOMA transmission states. If PU link $i, i \in \mathcal{P}$ uses SU $j, j \in \mathcal{S}$ as a relay, the utility of PU_i is defined as the throughput with this partnership, $U_{i,j}^{p} = C_{bj} = C_{ji}$. Here, we assume that the SAP_j relay should forward all the data received from the incumbent BS

to the cooperating PU_i, that is, satisfying the flow conservation constraint $C_{bj} = C_{ji}$, because the primary data has higher priority. The utility of SAP relay j is the sum of the throughput that it receives and transmits its own data on the subchannel leased from PU link i, $U_{i,j}^{s} = \sum_{k\in\mathcal{S}_j}(w_u C_{kj} + w_d C_{jk})$ where w_u and w_d, $0 \le w_u, w_d \le 1$, are the weight factors that are put on the SU uplink and SU downlink transmissions, respectively.

In the case of direct transmission from the BS to PU_i without cooperative relaying, the utility of PU_i is $U_i^{dir} = C_{bi,dir}$. The utility of SAP_j without cooperation is $U_j^{dir} = C_{kj} = C_{jk} = 0$ because the SAP does not have spectrum to transmit without cooperation. A PU_i selects a SAP_j as the cooperative relay only when its utility through the relay is greater than that of the direct transmission, i.e. $U_{i,j}^{p} = C_{bj} = C_{ji} > U_i^{dir} = C_{bi,dir}$. For a SU, the requirement is that its utility with cooperation should be greater than zero, i.e. $U_{i,j}^{s} = \sum_{k\in\mathcal{S}_j}(w_u C_{kj} + w_d C_{jk}) > U_j^{dir} = 0$. In addition, the transmit power allocation of the SAP should be subject to the SAP total power constraint: $\alpha_{ji}^2 + \sum_{k\in\mathcal{S}_j}\alpha_{jk}^2 \le P_{max}$, where P_{max} is the maximal total transmission power of SAP.

Let us first assume that PU_i has selected SAP_j as its cooperative relay. The relay selection optimization problem will be discussed in the next section. Then, the objective is to maximize the total utility $U_{i,j}^{p} + U_{i,j}^{s}$ by jointly determining the optimal subslot length δ and the power allocation of the SAP MMIMO-NOMA PU data relay and secondary data transmission to SU_k, $k \in \mathcal{S}_j$, subject to the above flow conservation and power constraints. The optimization problem can be formulated as

$$\max_{\delta, P_{ji}, P_{jk|k\in\mathcal{S}_j}} \{U_{i,j}^{p} + U_{i,j}^{s}\} = \max_{\delta, P_{ji}, P_{jk|k\in\mathcal{S}_j}} \{C_{ji} + \sum_{k\in\mathcal{S}_j}(w_u C_{kj} + w_d C_{jk})\} \tag{18}$$

$$\text{s.t. } C_{bj} = C_{ji} > C_{bi,dir},\ \sum_{k\in\mathcal{S}_j}\left(w_u C_{kj} + w_d C_{jk}\right) > 0,$$

$$\alpha_{ji}^2 + \sum_{k\in\mathcal{S}_j}\alpha_{jk}^2 \le P_{max}, \qquad 0 < \delta < 1$$

The above constrained optimization problem can be solved with gradient ascent algorithms [14].

3.2 Two-sided Matching

Next, we will focus on the cooperation and relay selection problem among multiple PU links and MMIMO-NOMA-empowered SAPs, and aim to optimize the utilities of all the entities with fairness. There exist competitions among the PUs, as well as among the SUs during relay selection and partner matching. The optimal strategy of an entity depends on the behaviors of other entities.

In practice, the SAP_j can estimate the channel vectors $\mathbf{h_{bj}}$ and $\mathbf{H_j} = [\mathbf{h_{1j}}, \ldots, \mathbf{h_{kj}}, \ldots \mathbf{h_{Kj}}]$ in its small cell and determine the power allocation and beamforming for its MMIMO-NOMA transmission. The PU_j estimates the channel coefficients h_{bi} and $\mathbf{h_{ij}}$. We assume that a common control channel is available for exchanging messages among the entities involved in cooperation. Then, BS, PUs and SAPs periodically exchange control messages within their transmission range, i.e. local neighborhood. All the achievable link rates can be then derived. We consider the scenario with-

out global information. The partnerships are formed through local information exchange among the BS, PUs and SAPs. Under this setting, we find our problem is best modeled using two-sided matching theory [9, 15, 16]. Based on this theory, we define the following concepts.

Definition 1: An entity is *individual rational*, if it will only cooperate with others when such a partnership improves its utility, i.e., $U_{i,j}^{p} > U_{i,dir}^{p}$, $\forall\, i \in \mathcal{P}$ and $U_{i,j}^{s} > 0$, $\forall\, j \in S$.

Definition 2: A *blocking pair* is a pair (PU_i, SAP_j) who both already have their respective partners $n(i)$ and $n(j)$, but prefer each other rather than their partners, i.e., $U_{i,n(i)}^{p} < U_{i,j}^{p}$ and $U_{n(j),j}^{s} < U_{i,j}^{s}$.

We can easily see that, if there exists a blocking pair, the entities involved have an incentive to break up from their existing partnership and form a new pair. Therefore, the current matching is unstable and not desirable. The definition of matching stability is given as follows.

Definition 3: A matching is stable if and only if every participating PU and SAP is individual rational and if there is no blocking pair in the network.

Based on the given definitions, our objective is to find a stable matching in the primary and secondary markets. It has been proven that a stable matching always exists for a two-sided market [15], and PUs and SAPs can find their partners using the following matching algorithm.

For each PU p ∈ $\mathcal{P}$:

```
Initialize the preference list by ranking the PU's utility in
partnership with each of available SAP relays, p.list();
p.partner ← free; end ← false;
while end != false
  if p.partner = free and p.plist != ∅ then
    s ← pop(p.plist());
    Send a "propose" message to s;
  if p receives an "accept" message from s then
    p.partner ← s;
  if p receives a "reject" message then
    delete the message sender from p.plist();
    If the message sender is the current partner of p then
      p.partner ← free;
Algorithm end when no message to issue and no response re-
ceived from SAPs.
```

For each SAP s ∈ $\mathcal{S}$:

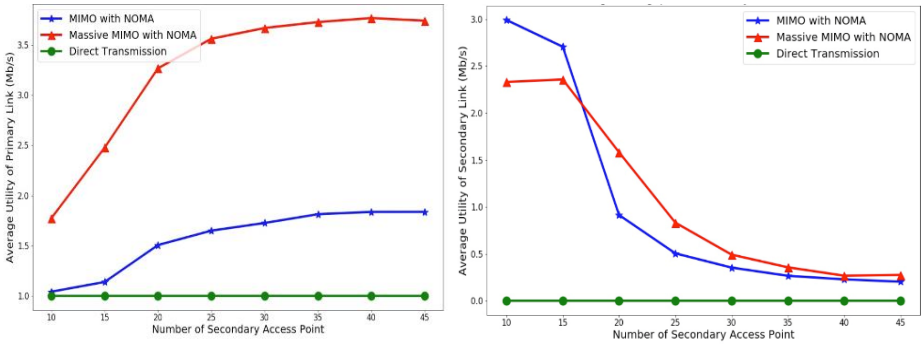
```
Initialize the preference list by ranking the SU's utility in
partnership with each of available primary links, s.list();
s.partner ← free; end ← false;
while end != false
  if s receives a "propose" message from p then
    if p ∉ s.list()then
        s sends a "reject" message to p;
    else
        s.partner ← p;
        s sends an "accept" message to p;
```

```
            for each PU p′ with a rank lower than p in s.list()
               s sends a "reject" message to p′;
               remove p′ from s.list();
    Algorithm end when no message received from PUs and no re-
    sponse to send.
```

The above algorithm is an extension of [16], which is a distributed version of the Gale–Shapley algorithm [15]. It can be proven that the algorithm finishes in $O(N_P + N_S)$ iterations and results in a stable matching that is optimal for the PUs [15], where N_P is the number of PUs and N_S is the number of SAP relays. Note that since the PUs represent the owners of the channel and can proactively lease the channel for higher utilities, the result of our mechanism is therefore desirable.

4 Evaluation Results

In this section, we evaluate the performance of the proposed MMIMO-NOMA-based cognitive cooperative relaying framework. We consider that N_P PUs are randomly located in a semicircle with a radius of 100 meters centered at the incumbent BS. Moreover, N_S MMIMO-NOMA SAP relays are randomly distributed within the same semicircle. Each SAP serves a small cell with 8 active SUs, 4 transmitters and 4 receivers, besides relaying data for the PU. The SUs are randomly placed in a circle centered at the SAP with a radius of 25 meters. A SU receives or sends the secondary data from or to the SAP. In addition, we assume that the subchannel bandwidth for a PU link is 1 MHz. The thermal noise level is set to be -100 dBm. The transmission power of the incumbent BS, P_{bs}, is set to be a value such that the average channel SNR of the PUs is 0 dB. The maximum transmission power of a SAP and a SU is set to be $1.0 \times P_{bs}$ and $0.5 \times P_{bs}$, respectively. We further assume that the incumbent BS and PUs are equipped with a single antenna. The SAPs are equipped with MMIMO and NOMA transceivers, and the SU has a single antenna, but has NOMA capability. As discussed before, the channel is modeled as a flat block-fading channel [10] and line-of-sight propagation between the sender and receiver with a path loss exponent of 3 and a small-scale Rayleigh fading component $\sigma = 1$.



(a) **(b)**

Fig. 3. (a) average utility of primary users, (b) average utility of secondary access points versus the number of secondary access points for difference schemes.

Figs. 3(a) and 3(b) illustrates the average utilities of PUs and SAP relays, respectively, versus the number of SAPs. The number of PU links, N_P is set to be 20. The "Direct Transmission" in the figures means that there is no cooperative SAP relaying, and the incumbent BS directly transmits its data to a PU. For the "MIMO with NOMA" scheme, the SAP relay has only two antenna elements. With massive MIMO and NOMA, the SAP equips with a 32-element MMIMO antenna array. Fig. 3(a) shows the average PU utility for the relaying schemes improves as the number of SAPs increases because more SAPs result in more opportunities for the PU links to find suitable cooperative relays. Fig. 3(b) shows that the SU utility for the relaying schemes decreases as the number of SAPs increases because more SAPs compete to access the limited spectrum resource. We can see from the figures that by exploiting cooperative SAP relaying with massive MIMO and NOMA, the proposed scheme significantly outperforms the direct transmission and MIMO-NOMA relaying schemes in terms of PU's utility. The MIMO-NOMA relaying scheme sometimes achieves a little better SAP utility than the MMIMO-NOMA relaying scheme, especially in the case of fewer SAP relays, because the SAP relays may not be at a good location and need to greedily allocate more power to transmit the PU data, but leave less power for SU data transmission.

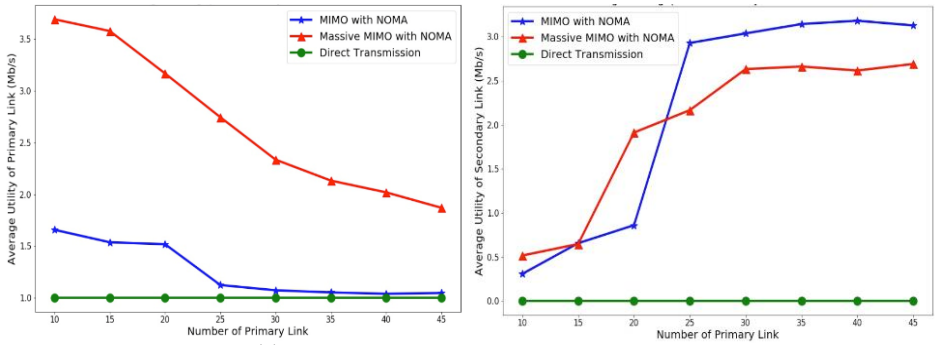


Fig. 4. (a) average utility of primary users, (b) average utility of secondary access points versus the number primary users for difference schemes.

Figs. 4(a) and 4(b) shows the utilities of PUs and SAPs versus the number of PU links, respectively, when the number of SAPs, N_s is 20. In Fig. 4(a), the average PU utility decreases for the relaying schemes as the number of PUs increases because more PU links compete for good SAP relays, and some of PUs may not be able to find suitable relays. In Fig. 4(b), the SAP utility improves with more PUs because a SAP is more likely to be selected as a relay for a PU link and access the PU's spectrum for its own data transmission. The MMIMO-NOMA based cooperative relaying scheme achieves much higher PU utility than the baselines. Due to the greedy allocation of the SAP transmit power to maximize the PU utility, the SU utility of the MMIMO-NOMA relaying scheme is less than that of the MIMO-NOMA relaying scheme in some cases. The results validate that our MMIMO-NOMA cooperative relaying framework can achieve win-win gains for both PUs and SUs.

5 Conclusions

In this paper, we present a novel framework that enables multiple PUs and multiple MMIMO-NOMA empowered SAPs to cooperate in traffic relaying and dynamic spectrum sharing. By leveraging the MMIMO and NOMA capabilities, SAPs help relay traffic for PUs while concurrently accessing the PUs' spectrum to transmit their own data. The optimization algorithms for the SAP relay selection and data transmission are proposed and analyzed. Evaluation results show that both PUs and SAPs can benefit from this proposed framework.

References

1. L. Lu, G. Li, A. Swindlehurst, A. Ashikhmin and R. Zhang: An overview of massive MIMO: Benefits and challenges. IEEE J. of Sel. Topics in Signal Processing, 8(5), 742–758 (2014).
2. J Hoydis, S ten Brink, M Debbah: Massive MIMO in the UL/DL of cellular networks: how many antennas do we need: IEEE J.Sel. Areas Commun. 31(2), 160–171 (2013).
3. 3GPP TS 38.201, ver. 15.0.0, Rel. 15, 5G; NR Physical layer General description (2018).
4. 3GPP TS 38.214, ver. 15.4.0, Rel. 15, 5G; NR; Physical layer procedures for data, (2018).
5. Ding, Zhiguo, Linglong Dai, Robert Schober, and H. Vincent Poor. "NOMA meets finite resolution analog beamforming in massive MIMO and millimeter-wave networks." IEEE Communications Letters, 21(8), 1879-1882 (2017).
6. Ding, Zhiguo, Yuanwei Liu, Jinho Choi, Qi Sun, Maged Elkashlan, I. Chih-Lin, and H. Vincent Poor. "Application of non-orthogonal multiple access in LTE and 5G networks." IEEE Communications Magazine 55(2), 185-191 (2017).
7. Morgado, A., Saidul Huq, K.M., Mumtaz, S., Rodriguez, J., A survey of 5G technologies: regulatory, standardization and industrial perspectives, Digital Communications and Networks Journal, 4(2), 87-97 (2018).
8. M. Song, C. Xin, Y. Zhao, and X. Chen, "Dynamic spectrum access: from cognitive radio to network radio," IEEE Wireless Comm., pp. 23-29 (2012).
9. S. Hua, H. Liu, X. Zhuo, M. Wu, and S. Panwar, "Exploiting Multiple Antennas in Cooperative Cognitive Radio Networks," IEEE Transactions on Vehicular Technology, 63(7), 3318-3330 (2013).
10. J. Zhang and Q. Zhang, "Stackelberg game for utility-based cooperative cognitive radio networks," in Proc. ACM MOBIHOC (2010).
11. E. Biglieri, et al.: MIMO Wireless Communications. Cambridge University Press (2007).
12. X. Gao, L. Dai, Y. Sun, S. Han, and C.-L. I: Machine learning inspired energy-efficient hybrid precoding for mmWave massive MIMO systems, in Proc. IEEE Int. Conf. on Commun., Paris, France (2017).
13. H. Ngo, E. Larsson and T. Marzetta: Energy and spectral efficiency of very large multiuser MIMO systems. IEEE Trans. Commun., 61, 1436–1449 (2013).
14. C. Papadimitriou and K. Steiglitz, Combinatorial Optimization, Algorithms and Complexity, Dover Publication, Inc., New York (2014).
15. A. E. Roth and M. A. O. Sotomayor, Two-Sided Matching: A Study in Game-Theoretic Modeling and Analysis. Cambridge, U.K.: Cambridge Univ. Press (1992).
16. I. Brito and P. Meseguer, "Distributed stable matching problems," Proc. Principles Pract. Constraint Programm., pp. 675–679 (2005).