A Practical Downlink NOMA Scheme for Wireless LANs

Pedram Kheirkhah Sangdeh, Hossein Pirayesh, Qiben Yan, Kai Zeng, Wenjing Lou, and Huacheng Zeng

***Abstract*—Non-orthogonal multiple access (NOMA) has emerged as a new multiple access paradigm for wireless networks. Although many results have been produced for NOMA, most of them are limited to theoretical exploration and performance analysis in cellular networks. Very limited progress has been made so far in the design of practical NOMA schemes for wireless local area networks (WLANs). In this paper, we propose a practical downlink NOMA scheme for WLANs and evaluate its performance in real-world wireless environments. Our NOMA scheme has three key components: precoder design, user grouping, and successive interference cancellation (SIC). On the transmitter side, we first formulate the precoding design problem as an optimization problem and then devise an efficient algorithm to construct precoders for downlink NOMA transmissions. We further propose a lightweight user grouping algorithm to ensure the success of SIC at the receivers. On the receiver side, we propose a new SIC method to decode the desired signal in the presence of strong interference. In contrast to existing SIC methods, our SIC method does not require channel estimation to decode the signals, thereby improving its resilience to interference. We have built a prototype of the proposed NOMA scheme on a wireless testbed. Experimental results show that, compared to orthogonal multiple access (OMA), the proposed NOMA scheme can significantly improve the weak user's date rate (93% on average) and considerably improve WLAN's weighted sum rate (36% on average).**

***Index Terms*—NOMA, successive interference cancellation (SIC), WLAN, experimentation**

I. Introduction

Multiple access is a crucial mechanism for wireless network infrastructure to serve multiple users. Orthogonal multiple access (OMA) techniques (e.g., time-division multiple access (TDMA) and frequency-division multiple access (FDMA)), albeit easy to implement, are incapable of approaching network capacity limit due to their exclusivity in resource allocation. This issue becomes particularly acute for networks with strict user fairness requirements. Non-orthogonal multiple access (NOMA) has recently emerged as a new multiple access paradigm for infrastructure-based wireless networks. Since its inception, NOMA has attracted a large amount of research attention and has been widely regarded as a promising candidate for radio access technologies (RAT) for 5G networks and beyond. In contrast to OMA, NOMA allows multiple users to utilize the same spectrum band for signal transmissions at the same time and, therefore, offers many advantages such as improving spectral efficiency, enhancing resource allocation flexibility, reducing scheduling latency, increasing cell-edge throughput, and enabling massive connectivity.

Recognizing its great potentials, power-domain NOMA has been studied in a variety of network settings in an increasingly sophisticated form, such as power allocation in single-input single-output (SISO) networks [1]–[3], precoder design in multi-input single-output (MISO) networks [4]–[7], and privacy protection [8]–[10]. Although a considerable amount of research efforts have been made on the study of NOMA, most of them are limited to theoretical exploration and performance analysis in cellular networks. Very limited progress has been made so far in the development of practical NOMA schemes and experimental validation of NOMA in real wireless network settings. This stagnation reflects the challenges in the design of practical NOMA schemes and the engineering issues related to their implementations, such as channel acquisition and precoding on the transmitter side and successive interference cancellation (SIC) realization on the receiver side.

In this paper, we aim to make a concrete step forward to bridge this gap by proposing a practical downlink NOMA scheme for WLANs and evaluating its performance on a wireless testbed. We consider an access point (AP) that has one or multiple antennas and a set of widely distributed users that have one antenna each. In such a network setting, we first examine the precoder design problem at the AP for downlink NOMA transmissions. We formulate the precoder design problem as an optimization problem, which inevitably includes non-convex constraints due to the intrinsic complication of the problem. To solve this problem, we employ a minorization-majorization (MM) approach for convexification of constraints. Based on the convexification results, we further develop an iterative algorithm to solve the precoder design problem.

Based on the solution to the precoder design problem, we develop a downlink NOMA scheme to enable concurrent data transmissions from an AP to multiple users. Our NOMA scheme features a lightweight user grouping strategy and a new SIC method. Specifically, on the transmitter (AP) side, we develop a heuristic algorithm to group the users for downlink NOMA transmission; on the receiver (user) side, we propose a robust SIC algorithm for interference subtraction and signal detection. In contrast to existing SIC methods [11], which first estimate the channels and then use the estimated channels to decode the signal/interference sequentially, our proposed SIC method does not require channel knowledge for interference subtraction and signal detection. Instead, it directly uses the reference signals (the precoded preamble

P. Kheirkhah Sangdeh, H. Pirayesh, and H. Zeng are with the Department of Electrical and Computer Engineering, University of Louisville, Louisville, KY 40292 (*Corresponding author*: H. Zeng). Q. Yan is with the Department of Computer Science and Engineering, Michigan State University, East Lansing, MI 48824. K. Zeng is with the Department of Electrical and Computer Engineering, George Mason University, Fairfax, VA 22030. W. Lou is with the Department of Computer Science, Virginia Tech, Falls Church, VA 22043.
This work is supported in part by the NSF grants CNS-1717840 and CNS-1846105.

in a frame) to compute the detection filters, which are used for interference subtraction and signal detection. As channel estimation is vulnerable to interference, the removal of channel estimation in the SIC procedure improves the performance and reliability of signal detection in our NOMA scheme.

We have built a prototype of the proposed NOMA scheme on a GNURadio-USRP2 wireless testbed using IEEE 802.11 legacy parameters and conducted extensive experiments in indoor office wireless environments to evaluate its performance in comparison with a conventional TMDA-based OMA scheme. We consider the following three network settings: (i) the AP has one antenna and it serves two users; (ii) the AP has two antennas and it serves two users; and (iii) the AP has two antennas and it serves three users. Our experimental results show that, compared to OMA, the proposed NOMA scheme can significantly improve the data rate of the weak user and considerably improve the weighted sum rate of all users. Specifically, for the cases that we have examined, the average improvement of data rate of the weak users is about 93.1%, and the average improvement of weighted sum rate of all users is about 36.1%. Moreover, our experimental results show that, on average over all the cases that we have considered, our proposed SIC method outperforms the conventional SIC method (least-squares channel estimation and zero-forcing signal detection) by 13.4% for the AP's weighted sum rate and 39.6% for the data rate of users performing SIC.

The remainder of this paper is organized as follows. Section II surveys the related work. In Section III, we describe the problem to be solved and offer a premier of NOMA. In Section IV, we study the precoder design problem for NOMA and devise an algorithm to solve this problem. In Section V, we propose an NOMA scheme for WLANs. Section VI delineates our implementation details and presents the experimental results. Finally, Section VII concludes this paper.

II. Related work

Since its inception, NOMA has been studied in an increasingly sophisticated form for cellular networks. Given that this work studies power-domain NOMA for the downlink of wireless networks, we focus our literature review on this specific area.

Power Allocation for NOMA. Power allocation for NOMA has been well studied in cellular networks where each node has a single antenna. These research efforts mainly focus on the power allocation strategies for NOMA under different performance considerations, such as user fairness [12], [13], outage probability [14], [15], and achievable throughput [16], [17]. This research line was then expanded to joint optimization of power allocation and subcarrier assignment for NOMA in OFDMA networks. These research efforts have produced many results, such as maximizing sum rate subject to the power constraints [1]–[3], minimizing the power consumption subject to SIC and rate requirements [18], and developing tractable algorithms [19].

Precoder Design for NOMA. When the base station (BS) has multiple antennas, the power allocation problem in downlink NOMA is escalated to precoder design problem as the power allocation and beam steering operations are tightly coupled. The precoder design at the BS needs to jointly optimize NOMA's power allocation and MIMO's beam steering. In the literature, precoder design has been studied toward different objectives, such as maximizing network throughput [4], [5], maximizing transmitter's energy efficiency [6], [7], and preserving users' signal privacy [8]–[10]. In what follows, we discuss the papers that are mostly relevant to our work.

In [4], the precoder design problem has been studied for NOMA to maximize sum rate, subject to the SINR constraints in the SIC process at all the users. A non-convex optimization problem was formulated and an iterative algorithm was developed to pursue a feasible solution. In [5], the precoder design problem has been studied to maximize the sum rate of a sophisticated hybrid network where an unmanned aerial vehicle and a BS serve a set of ground users. Precoder design aimed to nullify the cross-network interference or maintain the interference below a certain threshold.

The precoder design problem has also been studied for security enhancement and privacy preservation for power-domain NOMA. In [20], NOMA was studied under eavesdropping attacks, and artificial jamming approach was studied to combat the attacks. A non-convex optimization problem was formulated to maximize the artificial jamming power and, similar to [4], an iterative algorithm was developed to solve the problem. In [9], precoders were designed to ensure the privacy of a particular user. Specifically, precoders were designed to ensure that the private user's signal is of the weakest strength at all the users except itself. By doing so, none of the users are capable of decoding the private user's message.

As we shall see, our mathematical formulation of the precoder design problem is different from those existing ones. It features practical considerations in the design of precoders for downlink NOMA.

User Grouping for NOMA. User grouping is another key component of NOMA. In [21], the impact of user grouping on the performance of NOMA was studied. It shows that the throughput gain of NOMA (over OMA) becomes more significant when the channel strengths of the users in a group increases. However, reaching the optimal user grouping solution demands an exhaustive search. In [22] and [23], it was shown that the computational complexity of exhaustive-search-based user grouping algorithm can be relaxed by pruning the search space. Greedy grouping algorithms (e.g., [24]) and matching-based grouping algorithms (e.g., [25]) were proposed to reach a near-optimal solution. To further reduce the computational complexity, [26] proposed a random grouping algorithm, which needs a very low computation. However, this random algorithm cannot fully exploit the throughput gain of NOMA. The user grouping algorithm in our work is a lightweight heuristic algorithm, and it is amenable to practical implementation.

Experimental Validation of NOMA. While there is a large body of theoretical work on NOMA, experimental validation of NOMA in real wireless environments remains limited. Some pioneering work can be found in [27]–[30]. Our work differs from these research efforts in the following two aspects. First, these research efforts study NOMA in cellular

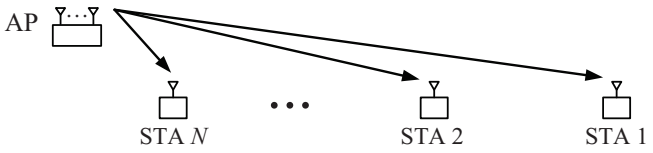


Fig. 1: Downlink data transmission in a WLAN.

networks, while our work focuses on NOMA for WLANs. Cellular networks and WLANs have significant differences in many aspects, including frame format, transmission pattern, transmit power, and receiver sensitivity. The results of NOMA in cellular networks cannot be directly applied to WLANs. Second, existing experimental efforts primarily investigated the gain of NOMA over OMA with respect to different system parameters and did not take into account precoder design for the performance optimization of NOMA. Our work considers both precoder optimization and NOMA implementation in WLANs.

III. Problem Description

We consider a WLAN as shown in Fig. 1, which comprises an AP and a set of user devices (a.k.a. stations, STAs, or users for simplicity). The AP has one or more antennas, and each station has a single antenna. Denote M as the number of antennas on the AP. Denote $\mathcal{N}$ as the set of stations, with N being its cardinality ($N = |\mathcal{N}|$). In this network, we assume that the signal from the AP to the stations experience significantly different path losses. That is, the signal received by STA i is much stronger than the signal received by STA $i-1$, for $2 \leq i \leq N$. This assumption can be fulfilled through a user selection/scheduling algorithm at the upper layer.

A Premier of NOMA. In power domain, NOMA takes advantage of the power difference between the interference and desired signals to mitigate interference and decode the desired signal at a receiver. SIC is typically used at the receivers for interference mitigation and signal decoding [11]. To illustrate the original concept behind power-domain NOMA, let us consider the network in Fig. 2 as an example. In this network, a single-antenna AP serves three stations standing far from each other. The AP sends superimposition of three signals to all stations: signal s_1 for STA 1, signal s_2 for STA 2, and signal s_3 for STA 3, with a proper power allocation for the these signals. At the stations, the received signals have significantly different strengths as illustrated in Fig. 2. The difference of signal strengths makes it possible for the stations to perform SIC. At STA 1, since the undesired signals s_2 and s_3 are relatively weak due to the large path loss, the desired signal s_1 can be easily decoded by treating interference (s_2 and s_3) as noise. At STA 2, the strongest undesired signal s_1 can be first decoded and subtracted from what received. For the resulting signals, the desired signal s_2 can be easily decoded by treating s_3 as noise. At STA 3, the strong undesired signal s_1 and s_2 can be first decoded and removed in a successive manner. After that, the desired signal s_3 can be decoded in a conventional way.

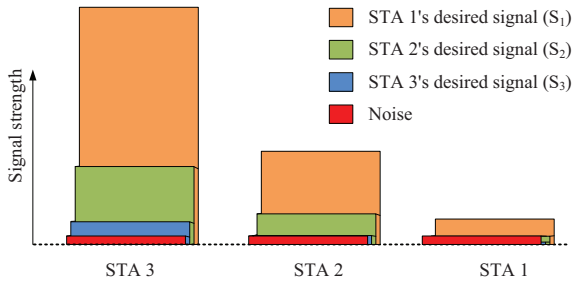


Fig. 2: Illustration of NOMA in downlink data transmission in a WLAN for $N = 3$.

As shown in this example, NOMA can enable concurrent data transmissions for STA 2 and STA 3 without causing much performance degradation for STA 1. Such a multi-user communication approach provides the AP with a new level of flexibility for its resource allocation and user scheduling, which can be leveraged for network throughput maximization.

Design Objectives. We consider the WLAN as shown in Fig. 1. We aim at developing a practical NOMA scheme to maximize the weighted sum rate of the users. For fairness, weak users have larger weights while strong users have relatively small weights. To do so, several questions remain open and needed to be addressed: (i) how to design the precoder for each data stream at the AP is a non-trivial task. When the AP has one antenna, the precoders degrade to complex coefficients, which represent the power allocation at the AP. When the AP has multiple antennas, the precoders determine not only the power allocation but also beam steering at the AP. The design of the optimal precoders at the AP involves both signal-to-interference-and-noise (SINR) and interference-to-signal-and-noise (ISNR) constraints at each station, making it challenging to reach the optimality. It is noteworthy that we design a unique precoder for each individual data stream in order to pursue performance optimality. Compared to the approach that separates the power allocation and beam-steering design, our joint design promises a better possible performance, especially for the networks with a small number of antennas and a small number of users. (ii) How to design a practical scheme to enable downlink NOMA in WLANs is another challenging problem. To support downlink NOMA, the AP needs the knowledge of channel state information (CSI) to compute the precoders; each station needs to perform signal detection in the face of inter-user interference. All these tasks require a sophisticated design of protocols and algorithms that are amenable to practical implementation.

IV. Precoder Design for Downlink NOMA

In this section, we first formulate the precoder design problem in downlink NOMA transmission of WLANs. Then, we convexify the non-convex constraints and propose an iterative algorithm to pursue a feasible solution. Finally, we offer discussions on the proposed algorithm.

A. Mathematical Formulation

Consider the downlink data transmission in the WLAN shown in Fig. 1. Denote $\mathbf{h}_i \in \mathbb{C}^{1\times M}$ as the channel from

the AP to STA i, which includes the effects of path loss, shadow fading, and fast fading. Owing to the large difference in path losses, we assume that $\|\mathbf{h}_1\| \leq \|\mathbf{h}_2\| \leq \ldots \leq \|\mathbf{h}_N\|$. At the AP, denote s_i as the signal intended for STA i, with $\mathbb{E}(|s_i|^2) = 1$; denote $\mathbf{v}_i \in \mathbb{C}^{M\times 1}$ as the precoding vector of this signal. The transmit signals at the AP, which is denoted by $\mathbf{x}$, can be written as $\mathbf{x} = \sum_{j\in\mathcal{N}} \mathbf{v}_j s_j$. Then, the received signal at STA $i \in \mathcal{N}$ can be written as:

$$y_i = \mathbf{h}_i \sum_{j=1}^{N} \mathbf{v}_j s_j + n_i, \qquad i \in \mathcal{N}\,. \tag{1}$$

where $n_i \sim \mathcal{CN}(0, \sigma_i^2)$ is additive white Gaussian noise.

Transmit Power Constraint. In practice, the transmit power of the AP is bounded by its maximum power budget, which we denote as P_{ap}. This constraint can be written as:

$$\sum_{i=1}^{N} \|\mathbf{v}_i\|^2 \leq P_{\text{ap}}\,. \tag{2}$$

SIC and SINR Constraints. At STA $i \in \{2, 3, \cdots, N\}$, we employ SIC to mitigate the strong interference $[s_1, s_2, \cdots, s_{i-1}]$ and decode the desired signal s_i by treating interference $[s_{i+1}, s_{i+2}, \cdots, s_N]$ as noise. Specifically, we first decode the undesired signal s_1 by treating signals $[s_2, s_3, \cdots, s_N]$ as noise. Based on the estimated signal $\hat{s}_1$, we can remove the effect of undesired signal s_1 and the resulting signal can be written as $y_i^{[1]} = \mathbf{h}_i \sum_{j=2}^{N} \mathbf{v}_j s_j + n_i$, where $y_i^{[1]}$ denotes the remaining signal after the first iteration of SIC. By the same token, we can continue to remove undesired signals $[s_2, s_3, \cdots, s_{i-1}]$ sequentially. After removing the undesired signals, we can decode the intended signal s_i by treating $[s_{i+1}, s_{i+2}, \cdots, s_N]$ as noise. Suppose that the SIC procedure is ideal. By denoting $\text{SINR}_{i,j}$ as the SINR in the jth iteration of SIC at STA i, we have

$$\text{SINR}_{i,j} = \frac{|\mathbf{h}_i \mathbf{v}_j|^2}{\sum_{k=j+1}^{N} |\mathbf{h}_i \mathbf{v}_k|^2 + \sigma_i^2}\,, \qquad i \in \mathcal{N}, 1 \leq j \leq i. \tag{3}$$

By defining $\gamma_{i,j}$ as a non-negative variable less than or equal to $\text{SINR}_{i,j}$, we have

$$\gamma_{i,j} \leq \frac{|\mathbf{h}_i \mathbf{v}_j|^2}{\sum_{k=j+1}^{N} |\mathbf{h}_i \mathbf{v}_k|^2 + \sigma_i^2}\,, \qquad i \in \mathcal{N}, 1 \leq j \leq i. \tag{4}$$

Data Rate Constraints. In the SIC procedure, STA i needs to decode signals $[s_1, s_2, \cdots, s_i]$ sequentially. When decoding signal s_j $(1 \leq j \leq i)$, we known that its SINR is greater than or equal to $\gamma_{i,j}$. To ensure that STA i can successfully decode s_j, the data rate of signal s_j is determined by this SINR value. Theoretically, the relationship between the maximum achievable data rate and the given SINR is governed by Shannon capacity. Denote r_j as the data rate from the AP to STA j in 1 Hz. Then, the achievable data rate constraints can be expressed as:

$$r_j \leq \log_2(1 + \gamma_{i,j}), \qquad i \in \mathcal{N}, 1 \leq j \leq i. \tag{5}$$

However, Shannon capacity is far from being reached by current WLANs' technologies. Therefore, it is highly inaccurate to characterize the relationship between the achievable data rate and the SINR in WLANs. In real wireless systems, adaptive modulation and coding scheme (MCS) is typically used to adjust the data rate based on the SINR value. Table I lists the MCS selection criteria that are specified in IEEE 802.11ac standard [31]. Fig. 3 shows the gap between Shannon capacity and the data rate achieved by adaptive MCS approach. It is evident that the gap is large. This indicates that Shannon capacity is not a good formula to compute the achievable data rate in real WLANs. To enhance the practicality of our results, we employ the adaptive MCS approach to calculate the achievable data rate for a given SINR value. However, their relation is expressed as a staircase function, which is non-convex. To ease our optimization problem, we approximate this non-convex region by the following linear constraints:

TABLE I: MCS specification in IEEE 802.11ac [31].

SINR (dB)	(inf,-5)	[-5,-10)	[-10,-13)	[-13,-16)	[-16,-19)	[-19,-22)	[-22,-25)	[-25,-27)	[-27,-30)	[-30,-32)	[-32,-inf)
Modulation	N/A	BPSK	QPSK	QPSK	16QAM	16QAM	64QAM	64QAM	64QAM	256QAM	256QAM
Coding rate	N/A	1/2	1/2	3/4	1/2	3/4	2/3	3/4	5/6	3/4	5/6
η	0	0.5	1	1.5	2	3	4	4.5	5	6	20/3
a_i	0.079	0.073	0.050	0.025	0.018	0.012	0.004	0.002	0.001	0.001	0
b_i	0	0.018	0.247	0.747	0.996	1.495	2.746	3.395	3.996	4.075	6.666

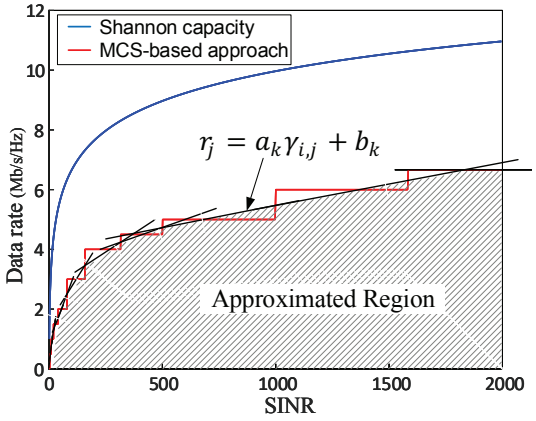


Fig. 3: The gap between Shannon capacity and the data rate achieved by MCS-based approach.

$$r_j \leq a_k \gamma_{i,j} + b_k, \qquad i \in \mathcal{N}, 1 \leq j \leq i, 1 \leq k \leq 11, \tag{6}$$

where a_k and b_k are constants given in Table I. We can see from Fig. 3 that, compared to (5), (6) is much more accurate to compute MCS-based achievable data rate in real WLANs. It is worth pointing out that the values of a_k and b_k in Table I were derived from the MCS specified in IEEE 802.11ac. If we want to apply this method to other networks such as IEEE 802.11ax, the values of a_k and b_k should be updated according to the MCS specified in corresponding standards.

Optimization Formulation. Based on the above constraints, we can formulate the NOMA problem as an optimization problem. Here, we consider the weighted sum rate as the objective function. Other objective functions (e.g., maximizing the minimum data rate) can be formulated in the same way. Denote w_j as the given weight for STA $j \in \mathcal{N}$. These weights are used to prioritize the service for the STAs and maintain the fairness among the STAs. Generally speaking, a STA with strong channel should be given a small weight, while a STA with weak channel should be given a large weight. Suppose that the STAs' weights are pre-defined. Then, the objective function can be written as: $\sum_{j\in\mathcal{N}} w_j r_j$. The optimization

problem, which we denote as OPT-NOMA, can be written as:

OPT-NOMA:

$$\max \quad \sum_{j\in\mathcal{N}} w_j r_j \tag{7a}$$

$$\text{s.t.} \quad r_j \le a_k\gamma_{i,j} + b_k, \quad i\in\mathcal{N}, 1\le j\le i, 1\le k\le 11; \tag{7b}$$

$$\gamma_{i,j} \le \frac{|\mathbf{h}_i\mathbf{v}_j|^2}{\sum\limits_{k=j+1}^{N} |\mathbf{h}_i\mathbf{v}_k|^2 + \sigma_i^2}, \quad i\in\mathcal{N}, 1\le j\le i; \tag{7c}$$

$$\sum_{i\in\mathcal{N}} \|\mathbf{v}_i\|^2 \le P_{\text{ap}}\,; \tag{7d}$$

where r_i, $\gamma_{i,j}$, and $\mathbf{v}_i$ are optimization variables; w_j, a_k, b_k, $\mathbf{h}_i$, and σ_i, and P_{ap} are given parameters. Note that we changed the equality in (4) to the inequality in (7c). It can be verified that this operation does not alter the optimal value.

Compared to many existing optimization formulations of NOMA (see, e.g., [4], [20]), one may notice that our formulation does not have explicit constraints to represent the decoding order requirements of SIC. Actually, constraints (7b) and (7c) in our formulation can ensure the success of SIC at every station. Consider STA i for example. Signal s_i is its desired signal and s_j $(1 \le j < i)$ is interference that should be removed. The combination of (7b) and (7c) ensures that interference s_j $(1 \le j < i)$ can be decoded and subtracted. It also ensures that the desired signal s_i can be decoded after subtracting interference s_j $(1 \le j < i)$. As such, our formulation can ensure the success of SIC, albeit without explicit constraints to enforce the decoding order requirements of SIC. Moreover, our formulation is in a simpler format and relatively easier to solve compared to the existing ones.

OPT-NOMA is a non-convex problem, which is NP-hard in general. There is no efficient algorithm that can find its optimal solution in polynomial time. In the rest of this section, we delve into the development of a tractable approach to pursue a suboptimal solution to OPT-NOMA via disciplinary convexification.

B. Constraint Relaxation via Disciplinary Convexification

In OPT-NOMA, (7a) and (7b) are linear and easy to handle by optimization solvers; however, (7c) and (7d) are not. In what follows, we focus on these two nonlinear constraints.

Constraint (7d): This constraint is convex, but in an indisciplined form. To transform it to a disciplined convex constraint, we rewrite it as a *Lorentz cone* [32, Ch. 2]:

$$\left\| \left[\mathbf{v}_1^{\mathsf{T}}, \mathbf{v}_2^{\mathsf{T}}, \ldots, \mathbf{v}_N^{\mathsf{T}}\right] \right\| \le \sqrt{P_{\text{ap}}}\,. \tag{8}$$

Constraint (7c): This constraint generates a set of $\frac{N(N+1)}{2}$ non-convex inequations and needs to be convexified into a disciplined form. To convexify (7c), we introduce an auxiliary variable $z_{i,j}$ and define $z_{i,j} \ge \sum_{k=j+1}^{N} |\mathbf{h}_i\mathbf{v}_k|^2 + \sigma_i^2$. Then, (7c) can be equivalently broken into the following two sets of constraints:

$$\gamma_{i,j} z_{i,j} \le |\mathbf{h}_i\mathbf{v}_j|^2\,, \qquad i\in\mathcal{N}, 1\le j\le i; \tag{9a}$$

$$\sum_{k=j+1}^{N} |\mathbf{h}_i\mathbf{v}_k|^2 \le z_{i,j} - \sigma_i^2\,, \qquad i\in\mathcal{N}, 1\le j\le i. \tag{9b}$$

To convexify (9), we first focus on (9a) and then on (9b). For (9a), it is a non-convex constraint because it has a quadratic term on its right hand side (RHS). To untangle this problem, we employ tangent point and Taylor expansion to approximate the quadratic term with an appropriate affine [20], [33]. To illustrate this idea, let us consider a differentiable convex function $f(\mathbf{v})$ for example. At any feasible point, say $\tilde{\mathbf{v}}$, a tangent function $g(\mathbf{v}, \tilde{\mathbf{v}})$ can be defined such that $f(\mathbf{v}) \ge g(\mathbf{v}, \tilde{\mathbf{v}})$, and the equality holds at $\mathbf{v} = \tilde{\mathbf{v}}$. The tangent function $g(\mathbf{v}, \tilde{\mathbf{v}})$ is a minorant of $f(\mathbf{v})$, and the solution to the approximated problem using tangent point $\tilde{\mathbf{v}}$ will majorize the minorant [33]. To further make this constraint disciplinary, the first-order Taylor expansion of $f(\mathbf{v})$ can be used as the tangent function since it removes the high-order nondisciplinary components of $f(\mathbf{v})$. Using the first-order Taylor expansion, the tangent function can be written as:

$$g(\mathbf{v}, \tilde{\mathbf{v}}) = f(\tilde{\mathbf{v}}) + \nabla f(\tilde{\mathbf{v}})^{\mathsf{H}}(\mathbf{v} - \tilde{\mathbf{v}}). \tag{10}$$

We apply this idea to the RHS of (9a). If $f_i(\mathbf{v}_j) = |\mathbf{h}_i\mathbf{v}_j|^2$ is defined, then we have $\nabla f_i(\mathbf{v}_j) = 2\mathbf{h}_i\mathbf{h}_i^{\mathsf{H}}\mathbf{v}_j$. The tangent function at $\tilde{\mathbf{v}}_j$ can be written as:

$$\begin{aligned} g_i(\mathbf{v}_j, \tilde{\mathbf{v}}_j) &= f_i(\tilde{\mathbf{v}}_j) + \nabla f_i(\tilde{\mathbf{v}}_j)^{\mathsf{H}}(\mathbf{v}_j - \tilde{\mathbf{v}}_j) \\ &= \mathbf{h}_i\tilde{\mathbf{v}}_j\tilde{\mathbf{v}}_j^{\mathsf{H}}\mathbf{h}_i^{\mathsf{H}} + 2\mathbf{h}_i\mathbf{h}_i^{\mathsf{H}}\tilde{\mathbf{v}}_j(\mathbf{v}_j - \tilde{\mathbf{v}}_j) \\ &= 2\mathbf{h}_i\tilde{\mathbf{v}}_j\mathbf{v}_j^{\mathsf{H}}\mathbf{h}_j^{\mathsf{H}} - \mathbf{h}_i\tilde{\mathbf{v}}_j\tilde{\mathbf{v}}_j^{\mathsf{H}}\mathbf{h}_i^{\mathsf{H}}. \end{aligned} \tag{11}$$

Given that both sides of original constraint (9a) are real values, we use $Re\big(g_i(\mathbf{v}_j, \tilde{\mathbf{v}}_j)\big)$ as the tangent function for $f_i(\mathbf{v}_j)$. Then, the RHS of (9a) can be approximated by

$$\begin{aligned} |\mathbf{h}_i\mathbf{v}_j|^2 &\approx Re\big(g_i(\mathbf{v}_j, \tilde{\mathbf{v}}_j)\big) \\ &= 2Re(\mathbf{h}_i\tilde{\mathbf{v}}_j\mathbf{v}_j^{\mathsf{H}}\mathbf{h}_j^{\mathsf{H}}) - \mathbf{h}_i\tilde{\mathbf{v}}_j\tilde{\mathbf{v}}_j^{\mathsf{H}}\mathbf{h}_i^{\mathsf{H}}\,, \end{aligned} \tag{12}$$

We apply the same method to convexify the left hand side (LHS) of (9a). To convexify the product of two variables, we define a bivariate function $f(\gamma, z) = \gamma z$. It is neither convex nor concave since its Hessian matrix is neither positive semidefinite nor negative semidefinite, and it also has a saddle point at $\gamma = z = 0$. However, this function can be expressed as summation of a convex function and a concave one, i.e., $f(\gamma, z) = f_1(\gamma, z) + f_2(\gamma, z)$, where $f_1(\gamma, z) = \frac{(\gamma+z)^2}{4}$ and $f_2(\gamma, z) = -\frac{(\gamma-z)^2}{4}$. To convexify $f(\gamma, z)$, it suffices to pursue the idea of using tangent function for its concave component. Since $f_2(\gamma, z)$ is a differentiable concave function, tangent function $g(\gamma, z, \tilde{\gamma}, \tilde{z})$ is a majorant of $f_2(\gamma, z)$. Indeed, $f_2(\gamma, z) \le g(\gamma, z, \tilde{\gamma}, \tilde{z})$. This majorant can be expressed as a tangent function at point $(\tilde{\gamma}, \tilde{z})$ as:

$$g(\gamma, z, \tilde{\gamma}, \tilde{z}) = \frac{1}{2}(\tilde{\gamma} - \tilde{z})(\gamma - \tilde{\gamma} + \tilde{z} - z) - \frac{1}{4}(\tilde{\gamma} - \tilde{z})^2\,. \tag{13}$$

Based on the tangent function in (13), we can approximate $f(\gamma, z) = \gamma z$ using $f_1(\gamma, z) + g(\gamma, z, \tilde{\gamma}, \tilde{z})$. Then, the LHS of (9a) can be approximated by

$$\gamma_{i,j} z_{i,j} \approx f_1(\gamma_{i,j}, z_{i,j}) + g(\gamma_{i,j}, z_{i,j}, \tilde{\gamma}_{i,j}, \tilde{z}_{i,j})$$

$$=\frac{1}{4}\left(\gamma_{i,j}+z_{i,j}\right)^2-\frac{1}{4}\left(\tilde{\gamma}_{i,j}-\tilde{z}_{i,j}\right)^2 - \frac{1}{2}\left(\tilde{\gamma}_{i,j}-\tilde{z}_{i,j}\right)\left(\gamma_{i,j}-\tilde{\gamma}_{i,j}+\tilde{z}_{i,j}-z_{i,j}\right). \quad (14)$$

Based on the relaxations in (12) and (14), the non-convex constraint (9a) can be approximated by the following convex constraint:

$$\begin{aligned}&\frac{1}{4}\left(\gamma_{i,j}+z_{i,j}\right)^2-\frac{1}{4}\left(\tilde{\gamma}_{i,j}-\tilde{z}_{i,j}\right)^2 \\ &-\frac{1}{2}\left(\tilde{\gamma}_{i,j}-\tilde{z}_{i,j}\right)\left(\gamma_{i,j}-\tilde{\gamma}_{i,j}+\tilde{z}_{i,j}-z_{i,j}\right) \\ \leq& 2Re(\mathbf{h}_i\tilde{\mathbf{v}}_j\mathbf{v}_j^{\mathsf{H}}\mathbf{h}_j^{\mathsf{H}})-\mathbf{h}_i\tilde{\mathbf{v}}_j\tilde{\mathbf{v}}_j^{\mathsf{H}}\mathbf{h}_i^{\mathsf{H}}, \quad i\in\mathcal{N}, 1\leq j\leq i. \quad (15)\end{aligned}$$

So far, we have convexified constraint (9a). Now, we focus on (9b), which is a *restricted hyperbolic* constraint. This constraint is convex but indisciplined. To make it disciplined, we first introduce an existing technique, and then apply it to transform (9b). Consider an indisciplined convex constraint $\theta^2\leq\alpha\beta$, $\alpha,\beta\in\mathbb{R}^+$ and $\theta\in\mathbb{R}$. Based on [34], we have:

$$\theta^2\leq\alpha\beta \Longleftrightarrow \left\|\left[\theta, \frac{(\alpha-\beta)}{2}\right]\right\| \leq \frac{(\alpha+\beta)}{2}, \quad (16)$$

where $\Longleftrightarrow$ means that the two sides are equivalent, and the RHS is a disciplined convex constraint. By taking advantage of this result, indisciplined convex constraint (9b) can be equivalently transformed to a disciplined convex constraint as follows:

$$\left\|\left[|\mathbf{h}_i\mathbf{v}_{j+1}|,\cdots,|\mathbf{h}_i\mathbf{v}_N|, \frac{\left(z_{i,j}-\sigma_i^2-1\right)}{2}\right]\right\| \leq \frac{\left(z_{i,j}-\sigma_i^2+1\right)}{2}, \quad i\in\mathcal{N}, 1\leq j\leq i, \quad (17)$$

The relaxed problem using the convexified constraints can be written as:

OPT-NOMA-RELAX:

$$\begin{aligned}\max \quad & \sum_{i\in\mathcal{N}} w_i r_i \\ \text{s.t.} \quad & \text{(7b), (8), (15), and (17).}\end{aligned}$$

OPT-NOMA-RELAX is a second-order cone programming (SOCP) problem, which can be solved in polynomial time by off-the-shelf optimization solvers such as CVX and CVXOPT [35].

C. Our Proposed Algorithm

Based on OPT-NOMA-RELAX, we propose an algorithm to solve the original problem OPT-NOMA. The proposed algorithm is an iterative algorithm. In each iteration, we solve OPT-NOMA-RELAX by taking the output results from the previous iteration as the input parameters (tangent points for convexification). The iterative algorithm terminates if the increase of the objective value is less than a pre-defined threshold (ϵ) or the number of iterations reaches a pre-defined bound (N_{iter}). For notational simplicity, when solving OPT-NOMA-RELAX in iteration l, we denote $\left[\underline{\tilde{\mathbf{v}}}^{[l-1]}, \underline{\tilde{\gamma}}^{[l-1]}, \underline{\tilde{\mathbf{z}}}^{[l-1]}\right]$ as the input parameters (the tangent points for convexification) and $\left[\underline{\mathbf{v}}^{[l]}, \underline{\gamma}^{[l]}, \underline{\mathbf{z}}^{[l]}, \underline{\mathbf{r}}^{[l]}\right]$ as the output results (the optimal solution to OPT-NOMA-RELAX). Alg. 1 presents our proposed algorithm.

Algorithm 1 Solving OPT-NOMA

Inputs: Network parameters $\mathbf{h}_i$, σ_i, $\mathcal{N}$, w_j, P_{ap}, a_k, b_k, and convergence threshold ϵ;
Outputs: A solution to OPT-NOMA $\left[\underline{\mathbf{v}}^*, \underline{\gamma}^*, \underline{\mathbf{z}}^*, \underline{\mathbf{r}}^*\right]$;

```
Compute initial tangent points [ṽ^[0], γ̃^[0], z̃^[0]] using Alg. 2;
Specify the max number of iterations (e.g., N_iter = 100);
for (l = 1; l ≤ N_iter; l++) do
    [v^[l], γ^[l], z^[l], r^[l]] ← solving OPT-NOMA-RELAX
    using [ṽ^[l-1], γ̃^[l-1], z̃^[l-1]] as its tangent points;
    [ṽ^[l], γ̃^[l], z̃^[l]] ← [v^[l], γ^[l], z^[l]]
    if ||r^[l] − r^[l-1]|| < ε then
        Break;
    end if
end for
[v^*, γ^*, z^*, r^*] ← [v^[l], γ^[l], z^[l], r^[l]];
```

Algorithm 2 Constructing an initial search point (The first tangential set for OPT-NOMA-RELAX)

Inputs: Network parameters $\mathbf{h}_i$, σ_i, $\mathcal{N}$, w_j, P_{ap}, and safety gap ε;
Outputs: An initial tangential set for OPT-NOMA-RELAX $\left[\underline{\tilde{\mathbf{v}}}^{[0]}, \underline{\tilde{\gamma}}^{[0]}, \underline{\tilde{\mathbf{z}}}^{[0]}\right]$;

```
Generate random values for [v̂_1^T, v̂_2^T, ··· , v̂_N^T]
for (i = 1; i ≤ N; i++) do
    v_i = sqrt( (1−ε)P_ap / Σ_{j=1}^N ||v̂_j||^2 ) v̂_i
end for
for (i ∈ N, 1 ≤ j ≤ i) do
    Calculate z_{i,j} = (1+ε)( σ_i^2 + Σ_{k=j+1}^N |h_i v_k|^2 )
    Calculate γ_{i,j} = (1−ε)|h_i v_j|^2 / z_{i,j}
end for
[ṽ^[0], γ̃^[0], z̃^[0]] ← [v, γ, z]
```

For such an iterative algorithm, an important question is how to construct an appropriate initial tangential set for the OPT-NOMA-RELAX problem in the first iteration. It is well known that the performance of many optimization problems is heavily reliant on their initial search points. A good initial point significantly accelerates the search process and therefore remarkably reduces the computational time of the algorithm. In light of this, we develop an algorithm to construct a good initial search point for the OPT-NOMA-RELAX problem in the first iteration. Alg. 2 shows our proposed algorithm. In this algorithm, we first randomly generate a set of vectors for $\underline{\tilde{\mathbf{v}}}^{[0]}$ and then normalize its amplitude to meet the power constraint. Upon initializing $\underline{\tilde{\mathbf{v}}}^{[0]}$, we then calculate $\underline{\tilde{\gamma}}^{[0]}$ and $\underline{\tilde{\mathbf{z}}}^{[0]}$ based on their respective constraints. In this process, a small number ε is used to ensure the strict feasibility of the tangential set and maximize its corresponding objective value.

D. Discussions on the Proposed Algorithm

Alg. 1 and Alg. 2 constitute our proposed algorithm to solve the optimization problem OPT-NOMA. We have the following remarks for the proposed algorithm:

Remark 1 (Feasibility). Our proposed algorithm yields a feasible solution to the original optimization problem OPT-NOMA. We pinpoint this by arguing that any feasible solution to OPT-NOMA-RELAX is also feasible to OPT-NOMA. Comparing the two optimization problems, we can see that the different constraints are (7c) in OPT-NOMA and (15) and (17) in OPT-NOMA-RELAX; other constraints are the same. Now, let us focus on these two constraints. The relaxation from (7c) to (15) and (17) is actually a minorization-majorization process [33]. That is, we replaced a concave function on the LHS with a tangent affine function and replaced a convex function on the RHS with a tangent affine function (see (12) and (14) respectively). Based on the properties of concave and convex functions, we know that (15) and (17) are more restrictive than (7c). In other words, a solution satisfying (15) and (17) certainly satisfies (7c). Therefore, we conclude that a solution feasible to OPT-NOMA-RELAX is also feasible to OPT-NOMA. According to Alg. 1 and Alg. 2, it is easy to see that the generated solution is feasible to OPT-NOMA-RELAX, which is also feasible to the original optimization problem OPT-NOMA.

Remark 2 (Convergence). Our proposed algorithm converges to a stationary point. Since the feasible region of OPT-NOMA-RELAX is expanding over the iterations in Alg. 1 [33], the value returned by the objective function is non-decreasing. Moreover, the solution yielded by each iteration (from solving OPT-NOMA-RELAX) is in the feasible region of OPT-NOMA. Because of the same objective function on both problems, Alg. 1 converges to a stationary point, which could be either a global or local optimal point.

Remark 3 (Computational Complexity). Our proposed algorithm (Alg. 1) has polynomial-time computational complexity. Alg. 1 is an iterative algorithm. In each iteration, its main work is solving the OPT-NOMA-RELAX problem (an SOCP problem). Given $M \leq N$ in NOMA, the complexity of each iteration is $O(N^6)$ [36]. Since the number of iterations in Alg. 1 is bounded by N_{iter}, the overall computational complexity of Alg. 1 is $O(N_{\text{iter}} \cdot N^6)$.

Remark 4 (Imperfect CSI). In the formulation of OPT-NOMA, we assumed perfect CSI for the design of precoders. However, in real systems, perfect CSI may not be available. In that case, we can use the measured (imperfect) CSI as the input to compute the precoders. Apparently, the imperfection of CSI may lead to a performance degradation.

V. A Downlink NOMA Scheme for WLANs

In this section, we propose a practical scheme based on the precoder design in the previous section to enable downlink NOMA transmissions in WLANs. We consider a WLAN as shown Fig. 4, which comprises an AP and a set of widely distributed stations (STAs). Denote $\mathcal{S}$ as the set of STAs in the network, with $S = |\mathcal{S}|$. The STAs are sorted in non-decreasing order based on their channel quality (i.e., $\|\mathbf{h}_i\|$, $i \in \mathcal{S}$). STA 1 is the weakest station and STA S is the strongest one. For such a network, we propose a downlink NOMA framework to support multi-user communications by leveraging the precoder optimization approach in Section IV.

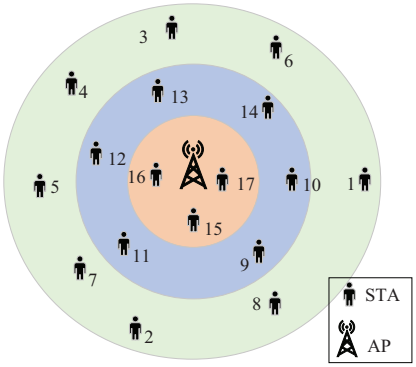


Fig. 4: An example of WLAN that has a set of widely distributed stations.

A. User Grouping at AP

We assume that the AP is responsible for user scheduling and grouping for the downlink transmissions. To perform user grouping, the AP needs to determine the number of stations in one group. Theoretical exploration of this problem requires an exhaustive search to identify the best grouping strategy that leads to the maximum network throughput. However, such an approach is overly complicated and not amenable to practical implementation. Therefore, we resort to a heuristic design for user grouping. In what follows, we first study the user grouping in a simple WLAN and then propose a heuristic algorithm for user grouping in a generic WLAN.

User Pairing in SISO Network. We consider a WLAN as shown in Fig. 4 and assume that each node (AP or STA/user) has a single antenna. We also assume that each group has two users in NOMA transmission for simplicity. Denote h_w and h_s as the channel coefficients of the weak and strong users in a group, respectively. Denote $p(h_w, h_s)$ as the normalized portion of AP's power allocated to the strong user's message. Based on the notion of NOMA, the AP's power allocation for NOMA transmission should have the following property: $p(h_w, h_s)$ is a non-increasing function with respect to $|h_s| / |h_w|$. Based on this property, we have the following proposition:

Proposition 1: Suppose that the objective is to maximize the weighted sum rate of all users and that round-robin scheduler is used for the paired users. Then, the best pairing strategy is $(i, S/2 + i)$*, and the worst pairing strategy is* $(i, S + 1 - i)$*, for* $1 \leq i \leq S/2$*, as illustrated in Fig. 5.*

The proof of Proposition 1 is given in Appendix. From Proposition 1, we have the following observations on user pairing: (i) it should try to avoid pairing two users with similar channel quality; and (ii) it should try to maintain a similar channel difference for user pairs.

User Grouping in MISO Network. Based on the above two observations, we propose a heuristic user grouping algorithm for a generic WLAN. For STA $i \in \mathcal{S}$, we define its channel

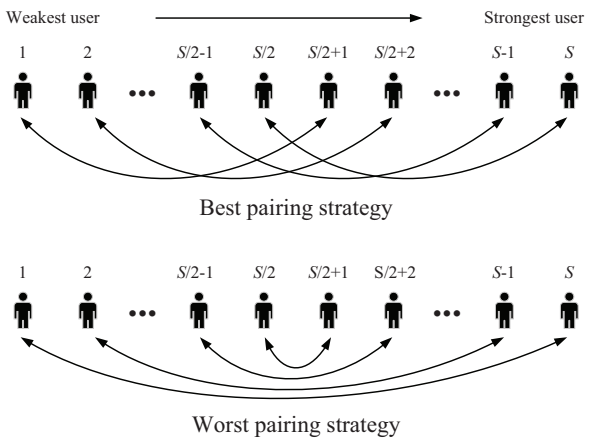


Fig. 5: Two pairing strategies in NOMA communications.

Algorithm 3 An algorithm for user grouping.

Inputs: The array of sorted STAs ($\mathcal{S} = \{1, 2, \cdots, S\}$) and each STA's channel ($\mathbf{h}_i$, $i \in \mathcal{S}$);
Outputs: The total number of groups (K) and the generated user groups $\mathcal{G}_1, \mathcal{G}_2, \cdots, \mathcal{G}_K$;

$\Delta q \leftarrow 10$;
$k \leftarrow 0$;
while ($\mathcal{S}$ is not empty) **do**
  k++;
  $\mathcal{G}_k = [\mathcal{S}(1)]$; // $\mathcal{S}(1)$ *is the 1st element of* $\mathcal{S}$
  $q\_value \leftarrow q(\mathcal{S}(1))$;
  for ($l = 2$; $l \leq size(\mathcal{S})$; l++) **do**
    if $q(\mathcal{S}(l)) \geq q\_value + \Delta q$ **then**
      $\mathcal{G}_k \leftarrow [\mathcal{G}_k \ \ \mathcal{S}(l)]$;
      $q\_value \leftarrow q(\mathcal{S}(l))$;
    end if
  end for
  Remove all elements in $\mathcal{G}_k$ from $\mathcal{S}$;
end while
$K \leftarrow k$;

quality indicator as $q(i) = 20\log_{10}(\|\mathbf{h}_i\|)$, where $\mathbf{h}_i$ is STA i's channel that includes path loss, shadow fading, and fast fading. Based on the channel quality indicator, we use the following rules to devise a user grouping algorithm: (i) the STAs in the same group should have at least Δq channel quality difference, where Δq represents the channel quality difference in decibel and should be adaptively set based on the network environment. In our experiments, extensive measurements of wireless channels in an office building show that the average channel quality difference of two users is about 9.3 dB. Based on this observation, we set $\Delta q = 10$ dB for the user grouping algorithm. (ii) one STA is associated with only one group. Per these two rules, we propose a greedy algorithm as shown in Alg. 3 for user grouping. Essentially, Alg. 3 is heuristic. We have the following remarks on it.

Remark 5 (Single STA in a Group): Based on our algorithm, it is apparent that there is no guarantee each group has more than one STA. If a group has only one STA, this means that NOMA is not needed, and OMA can be used for its transmission. Essentially, such a grouping algorithm requires a combination of NOMA and OMA at the PHY layer for data transmission.

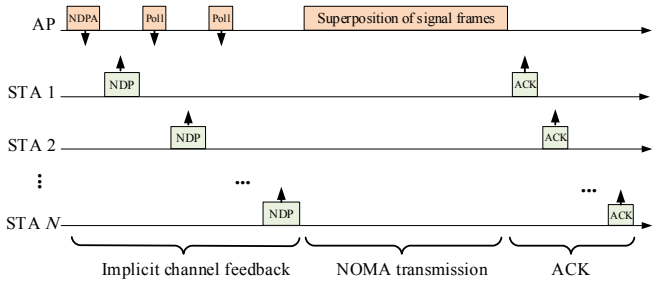


Fig. 6: A protocol for NOMA transmission in WLANs.

B. A MAC-Layer Protocol for NOMA

If a group includes multiple users (STAs), then NOMA is used to enable concurrent data transmission for the STAs. With a bit abuse of notation, we denote $\{1, 2, \cdots, N\}$ as the STAs in the group under consideration. Fig. 6 shows our proposed protocol for NOMA transmission. At high level, it comprises three steps: channel sounding, NOMA transmission, and acknowledgment. Since the acknowledgment step is straightforward, we focus our discussions on channel sounding and NOMA transmission.

Channel Sounding. To reduce the airtime overhead, we employ an implicit channel feedback mechanism in our protocol by leveraging the channel reciprocity. Specifically, the AP first broadcasts a Null Data Packet Announcement (NDPA) to inform the stations of channel sounding and NOMA transmission. Upon reception of the NDPA packet, the stations sequentially respond with a Null Data Packet (NDP) following the poll packets from the AP. The NDP includes the preamble (reference signals) enabling the AP to estimate the uplink channel. At the end of this step, the AP obtains the uplink channels between itself and all the intended stations. The obtained uplink channels will be converted to downlink channels through channel calibration.

In such an implicit channel feedback mechanism, three important problems need to be taken into consideration: (i) for the protocol in Fig. 6, the stations should use the same transmit power when transmitting the NDP (e.g., the maximum transmit power specified in the standards). Use of different transmit powers will confuse the AP about the channel quality between itself and the stations, thereby leading to a failure in the downlink NOMA transmission. (ii) Typically, the stations in a WLAN have the same noise power. In some extreme cases where the stations have different noise power, the stations need to feed their noise power back to the AP. This can be easily done by embedding the noise power information (only a real number) into the NDP when performing uplink channel sounding. (iii) To perform downlink NOMA transmission, the AP actually needs to know the downlink channels information. It is therefore imperative to infer the downlink channels based on the measured uplink channels in our protocol. When the AP has a single antenna, the difference between an uplink channel

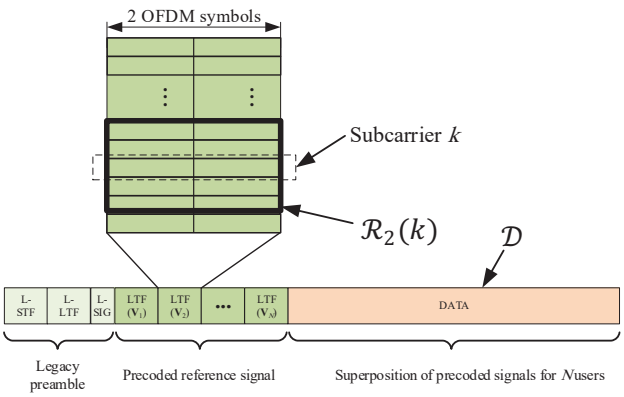


Fig. 7: Proposed frame structure for NOMA transmission.

and its corresponding downlink channel can be represented by a complex number in the mathematical channel model. Such a complex scalar does not affect the NOMA scheduling and transmission results. Therefore, the measured uplink channels can be equivalently treated as downlink channels. When the AP has multiple antennas, the difference between uplink and downlink channels is an array of complex numbers. To compensate the mismatch, a channel calibration procedure is needed at the AP. While there are many calibration methods, we employ the relative calibration method in [37]. This relative calibration method is an internal and standalone calibration method that can be done at the AP without any aid from the stations. In our experiment, we implicitly implement this calibration method to maintain the channel reciprocity.

NOMA Transmission and Frame Structure. After obtaining the downlink channel, the AP computes precoders using the proposed method in Section IV and selects a MCS for each STA in the scheduled group. Then, the AP performs downlink NOMA transmission as illustrated in Fig. 6. To perform downlink NOMA transmission, we propose a MU-MIMO-like frame structure as shown in Fig. 7. The proposed frame structure has three parts: (i) the legacy preamble part comprises a legacy short training field (L-STF), a legacy long training field (L-LTF), and a legacy signal (L-SIG) field. This part is designed for frame detection and time/frequency synchronization on the STA side. (ii) The reference signal part comprises N precoded LTFs, each of which has two identical OFDM symbols. This part is devised for signal detection on the STA side. (iii) The data (payload) part comprises a sequence of OFDM symbols, each of which is a superposition of precoded signals for N stations. In what follows, we propose a new SIC method to decode the desired signal at each STA.

C. PHY-Layer Signal Processing

At the PHY layer, multiple adjacent subcarriers are bonded together for data transmission in order to reduce the computational complexity. The rationale behind this operation is that the channels of adjacent subcarriers are typically similar. Hence, the bonding strategy does not cause much performance degradation but reduces the complexity significantly. Fig. 7 illustrates an example of bonding over five adjacent subcarriers. For the bonded subcarriers, we use the same precoder for power allocation and beam steering on the AP side and the same detection filter for signal recovery on the user side.

AP-Side Precoding. After computing the precoders for each user, we assemble a downlink NOMA transmission frame as shown in Fig. 7. The first part of the frame is fixed. The second part of the frame is computed based on the precoders. Specifically, for the jth LTF in this part, its frequency-domain data is generated by $\mathbf{x}(l,k) = \mathbf{v}_j(k)\bar{s}_j(l,k)$, where $\bar{s}_j(l,k)$ is a pre-defined reference signal. The third part is superposition of precoded data in the frequency domain. It is generated by $\mathbf{x}(l,k) = \sum_{j=1}^{N} \mathbf{v}_j(k)s_j(l,k)$. The generated signal vector $\mathbf{x}(l,k)$ is converted into time domain using IFFT operation. The resulting time-domain signal vector will be sent to RF chains for transmission.

User-Side SIC-based Signal Detection. Since each frame has the IEEE 802.11 legacy preamble, the users can perform frame detection, time synchronization, and frequency offset correction in the same way as conventional Wi-Fi devices do. Afterward, each user performs SIC to decode its desired signal. For ease of exposition, we denote l as the index of OFDM symbol in a frame and denote k as the index of subcarrier in the OFDM modulation. Then, the received signal at STA i can be written as:

$$y_i(l,k) = \sum_{j=1}^{N} \mathbf{h}_i(k)\mathbf{v}_j(k)s_j(l,k) + n_i(l,k). \quad (18)$$

To decode the desired signal at STA i, one approach is zero-forcing SIC (ZF-SIC). This approach decodes and subtracts the strongest signal sequentially, until its desired signal is obtained. When decoding the strongest signal, it simply treats other (non-strongest) signals as interference. When decoding s_j, it first estimates the compound channel by $\hat{h}_j(k) = \bar{y}_i(l,k)/\bar{s}_j(l,k)$, where $\bar{y}_i(l,k)$ and $\bar{s}_j(l,k)$ are the received and transmitted reference signals, respectively. Then, it uses the estimated channel to decode the strongest signal by letting $\hat{s}_j(l,k) = y_i(l,k)/\hat{h}_j(k)$, where $\hat{s}_j(l,k)$ is the estimated version of the strongest signal $s_j(l,k)$.

Although ZF-SIC is amenable to implementation, its performance is highly suboptimal. This is because it does not take into account the effect of noise and interference (non-strongest signals) in the course of its signal detection. To improve its performance, we may consider minimum mean square error SIC (MMSE-SIC), which takes into account noise and interference. In contrast to ZF-SIC, MMSE-SIC estimates the strongest signal as follows: $\hat{s}_j(l,k) = \hat{h}_j(k)^*[\hat{h}_j(k)\hat{h}_j(k)^* + \frac{1}{\rho}]^{-1}y_i(l,k)$, where ρ is the SINR. MMSE-SIC requires the knowledge of SINR, making it difficult to implement in practice.

To circumvent the above problems, we propose a new SIC scheme, which uses the reference signals to construct a detection filter directly. Specifically, at STA i, we decode signals $\{s_1(l,k), s_2(l,k), \cdots, s_i(l,k)\}$ in sequence. When decoding the strongest signal $s_j(l,k)$, we construct the detection filter as follows:

$$g_j(k) = \frac{\sum_{(l,k)\in\mathcal{R}_j(k)} y_i(l,k)s_j(l,k)^*}{\sum_{(l,k)\in\mathcal{R}_j(k)} y_i(l,k)y_i(l,k)^*}, \quad 1 \le j \le i, \quad (19)$$

Algorithm 4 The proposed SIC at STA i.

Inputs: Received signal $y_i(l,k)$, reference signals in the frame;
Outputs: Estimated signals in the data part of the frame, i.e., $\hat{s}_i(l,k)$ for $(l,k) \in \mathcal{D}$;
for $(j=1;\ j \le i;\ j$++$)$ **do**
  Compute decoding filter $g_j(k)$ using (19);
  Estimate current signal $\hat{s}_j(l,k)$ using (20);
  $\tilde{s}_j(l,k) \leftarrow$ QAM-based demodulation of $\hat{s}_j(l,k)$;
  $y_i(l,k) \leftarrow y_i(l,k) - \tilde{s}_j(l,k)/g_j(k)^*$;
end for

where $(\cdot)^*$ is the conjugate operator, and $\mathcal{R}_j(k)$ is the set of reference signals in the jth LTF on subcarrier k. Fig. 7 illustrates an example of $\mathcal{R}_j(k)$ when $j=2$. It can been seen that we use not only the reference signals on subcarrier k but also reference signals on its two neighboring subcarriers to construct detection filter $g_j(k)$. The rationale behind this design is that the summation over multiple subcarriers can reduce the effect of noise and interference (non-strongest signals). After calculating the detection filter, we estimate signal $s_j(l,k)$ in the data part of the frame as follows:

$$\hat{s}_j(l,k) = g_j(k)^* y_i(l,k), \quad 1 \le j \le i, \tag{20}$$

where $\hat{s}_j(l,k)$ is the estimated version of $s_j(l,k)$.

Based on (19) and (20), we present the proposed SIC algorithm in Alg. 4. Apparently, this algorithm does not require the estimated SINR. But, it can partially reduce the influence of noise and interference. This is important in SIC detection because all of non-strongest signals are considered as interference. Meanwhile, this SIC algorithm has a low complexity, and it is amenable to practical implementation. For its performance, we will show via experimental results that it considerably outperforms ZF-SIC.

VI. Performance Evaluation

In this section, we conduct experiments to evaluate the performance of the proposed NOMA scheme in real-world wireless environments.

A. Prototyping and Experimental Setup

Experimental Testbed. We have prototyped an AP and three users. The AP has been implemented using two USRP N210 devices and one laptop. Each user has been implemented using an USRP N210 device and one laptop. The USRP devices are used for radio signal transmission and reception, and the laptop is used for baseband signal processing. All the baseband signal processing is carried out by the laptop using Python and C++ in GNU Radio software package. The prototyped AP supports up to two antennas for signal transmission and reception, while users support one antenna.

On the AP, the relative calibration method in [37] was implemented to preserve the uplink/downlink channel reciprocity. This relative calibration method is a standalone calibration procedure that can be done by the AP without requiring involvement of the users.

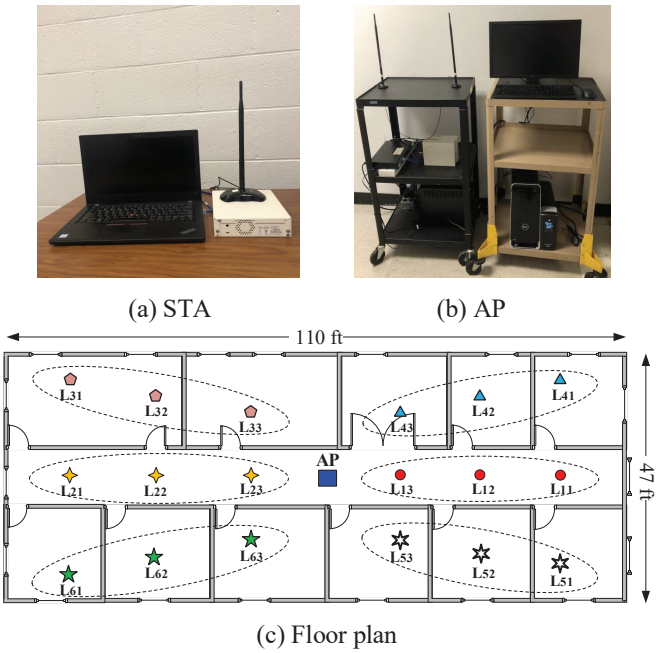
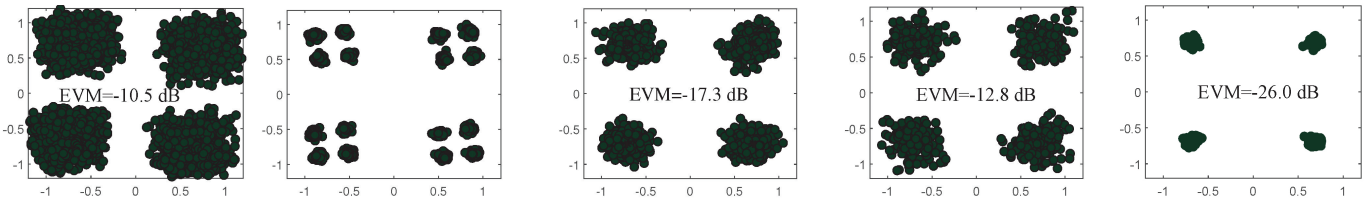


Fig. 8: Our NOMA testbed and floor plan.

Prototyping NOMA. We have implemented the NOMA protocol in Fig. 6 on the testbed. As shown in Fig. 6, the protocol first performs uplink channel sounding to obtain the uplink channels. Based on the channel knowledge, Alg. 1 is used to compute the precoders ($\mathbf{v}_i,\ i \in \mathcal{N}$) using a convex optimization solver such as CVX and CVXOPT [35] . In OPT-NOMA, we set $w_1 = 3$ for weak user, $w_2 = 2$ for middle user, and $w_3 = 1$ for strong user. These weights are just an example and other weights would also work. After computing the precoders, the AP sends a superimposition of the signals toward users using the frame structure depicted in Fig. 7. The users perform SIC to decode their desired signals. In our implementation, we use Schmidl-Cox algorithm for the timing and frequency synchronization at the receivers in both uplink and downlink transmissions. We use least-squares channel estimation at the AP in the uplink channel sounding. For the downlink transmissions, we use the precoded reference signals in the frame (see Fig. 7) to construct the channel equalization coefficient $g_j(k)$ and then use this coefficient for signal detection, as detailed in (19) and (20).

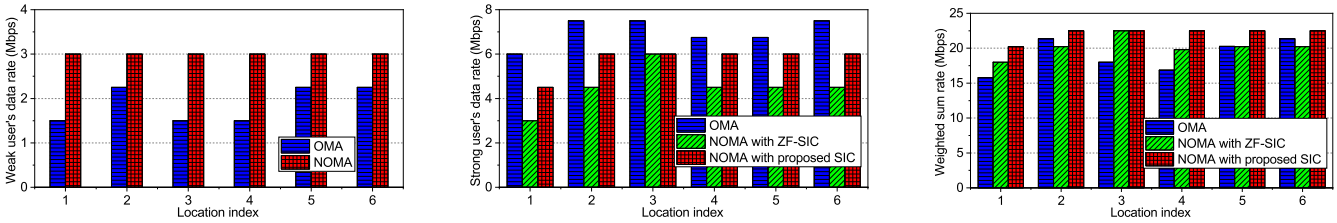
During this protocol, IEEE 802.11 legacy frame parameters are used for both uplink and downlink transmissions. That is, each OFDM symbol has 64 subcarriers in total; 48 subcarriers are used for data transmission; 4 subcarriers are used for pilot; and 12 subcarriers are null. The 52 valid subcarriers are bonded into two groups for the precoder design in NOMA. The length of cyclic prefix is 16. Due to the hardware limitation, we set the sampling rate to 5 Msps (to avoid the unflat circuit response from the CIC) and set the short interframe space (SIFS) to 2 seconds. Given the 5 MSps sampling rate, the time duration of each OFDM symbol is 16 µs. The data part of each frame consists of 20 OFDM symbols.

Experimental Setup: In our tests, the maximum transmit power for each node (AP or user) is set to 17 dBm. Fig. 8



(a) NOMA: decoded signal at weak user. (b) NOMA: decoded signal at strong user before SIC. (c) NOMA: decoded signal at strong user after SIC. (d) OMA: decoded signal at weak user. (e) OMA: decoded signal at strong user.

Fig. 9: Constellations of NOMA and OMA in downlink of WLAN where a single-antenna AP serves two single-antenna users.



(a) Weak user's data rate (b) Strong user's data rate (c) AP's weighted sum rate

Fig. 10: Performance comparison of NOMA and OMA in downlink transmission of a WLAN where a single-antenna AP serves two single-antenna users.

shows the floor plan of our test scenarios and the prototyped AP and STA. Regarding the floor plan, the AP is placed at a fixed location marked "AP". The three users are placed at one of the six different locations. Specifically, STA 1 is placed L_{k1}, STA 2 is placed L_{k2}, and STA 3 is placed L_{k3}, for $k = 1, 2, \cdots, 6$. It is noteworthy that the linear deployment of the three users in a group is for ease of explanation. In a rich scattering environment like an office building, such a deployment does not impose significant correlation among users' channels. Moreover, it is worth pointing out that our experimentation does not include the user grouping algorithm.

B. Performance Metrics and Benchmark

Performance Benchmark: We use OMA as the performance benchmark to evaluate the throughput gain of NOMA. In OMA, the round-robin scheduler is used at the AP. Specifically, the AP serves only one user in one time slot, and different users are scheduled in different time slots. When the AP has multiple antennas, the best antenna is selected for spatial diversity.

Performance Metrics: We evaluate the performance of the proposed NOMA scheme using the two metrics: Error Vector Magnitude (EVM) and data rate. (i) EVM is a metric widely used in WLANs. At STA j, its EVM is defined as $\text{EVM} = 10\log_{10}\left(\mathbb{E}_{l,k}\left[\,|\hat{s}_j(l,k) - s_j(l,k)|^2\,\right]/\mathbb{E}_{l,k}\left[\,|s_j(l,k)|^2\,\right]\right)$. (ii) The data rate is extrapolated based on the measured EVM at each user using the MCS specified in IEEE 802.11 [31]. Specifically, the data rate at STA j is calculated by

$$\text{NOMA:} \quad r_j = \frac{48}{80} \cdot b \cdot \eta(\text{EVM}), \tag{21a}$$

$$\text{OMA:} \quad r_j = \frac{1}{N} \cdot \frac{48}{80} \cdot b \cdot \eta(\text{EVM}), \tag{21b}$$

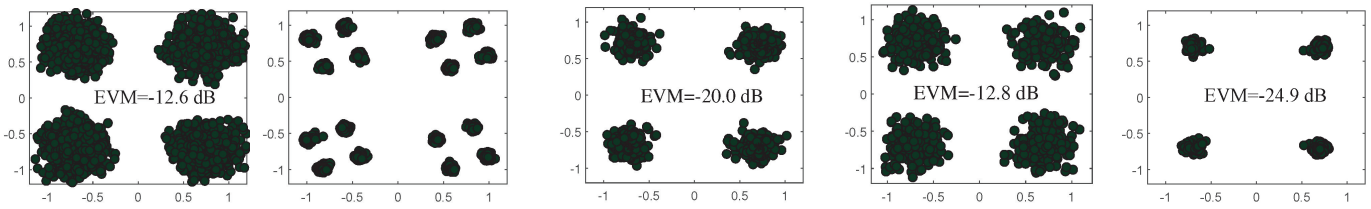
where N is the number of users served by the AP, 48 is the number of payload subcarriers, 80 is the number of samples in an OFDM symbol, b is the bandwidth (5 MHz), EVM is measured at the STA j when NOMA or OMA is used, and $\eta(\text{EVM})$ is the average number of bits carried by one subcarrier in an OFDM symbol and its value is given in Table I.

C. Experimental Results of (1×2)*-NOMA*

We first consider the case where the AP has one antenna and it serves two users (one weak user and one strong user). The weak user is placed at L_{k1} and the strong user is placed at L_{k3}, $k = 1, 2, \cdots, 6$.

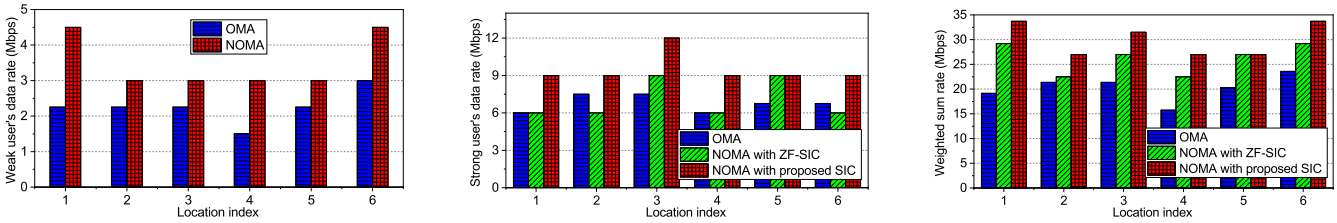
Case Study: We use location 4 ($k = 4$) as an example to examine NOMA. Fig. 9 presents the constellation of the decoded signals at the two users when NOMA and OMA are used, respectively. For the weak user, Fig. 9(a) and Fig. 9(d) show that NOMA has a small (2.3 dB) EVM degradation compared to OMA. Using (21), the data rate at the weak user is extrapolated to 3.0 Mbps in NOMA and 1.5 Mbps in OMA. This indicates that NOMA has a significant throughput gain for the weak user.

For the strong user, Fig. 9(b) and Fig. 9(c) show its decoded signals before and after SIC in NOMA, respectively. We can see that the strong user can achieve -17.3 dB EVM after SIC. Compared Fig. 9(c) with Fig. 9(e), we can see that the strong user has a considerable EVM degradation (about 9.4 dB) when NOMA is used. Using (21), the data rate achieved by the strong user is extrapolated to 6.0 Mbps in NOMA and 6.7 Mbps in OMA. Compared to OMA, our NOMA scheme slightly decreases the strong user's data rate. The reasons are two-fold. First, NOMA serves two users while OMA serves one user. Moreover, a higher weight is assigned for the weak user to maintain the fairness in NOMA transmission when we conduct the optimization (OPT-NOMA). Second, SIC in NOMA is not perfect due to the limited ADC resolution,



(a) NOMA: decoded signal at weak user. (b) NOMA: decoded signal at strong user before SIC. (c) NOMA: decoded signal at strong user after SIC. (d) OMA: decoded signal at weak user. (e) OMA: decoded signal at strong user.

Fig. 11: Constellations of NOMA and OMA in downlink of WLAN where a two-antenna AP serves two single-antenna users.



(a) Weak user's data rate (b) Strong user's data rate (c) AP's weighted sum rate

Fig. 12: Performance comparison of NOMA and OMA in downlink transmission of a WLAN where a two-antenna AP serves two single-antenna users.

circuit nonlinearity and distortion. The imperfection of SIC degrades the performance of the strong user.

From the AP's perspective, the weighted sum rate of the two users is 22.5 Mbps in our NOMA scheme and 16.9 Mbps in OMA. This means that our NOMA scheme has about 33.5% improvement over OMA in terms of weighted sum rate.

Results from All Locations: Fig. 10 presents the extrapolated data rate at each user when NOMA and OMA are used, respectively. Two SIC techniques are implemented for the strong user: ZF-SIC and our proposed SIC. Based on the results, we have the following observations: (i) for the weak user, our NOMA scheme has a 60.0% data rate gain compared to OMA, on average over all the locations that we tested. (ii) For the strong user, NOMA yields a 17.9% data rate degradation on average compared to OMA. This degradation can be attributed to the inter-user interference and the imperfections of SIC as explained above. (iii) For the AP, the proposed NOMA scheme outperforms OMA by 18.0% in terms of weighted sum rate. (iv) The proposed SIC scheme outperforms ZF-SIC by 27.8% for the stronger user's data rate. This throughput gain is from the summation operation in (19). Mathematically, this summation operation is equivalent to a low pass filter, which reduces the effects of noise and interference (weak signals) in each iteration of SIC.

D. Experimental Results of (2×2)*-NOMA*

We now consider the case where the AP has two antennas and it serves two users. The weak user is placed at L_{k1} and the strong user is placed at L_{k3}, $k = 1, 2, \cdots, 6$.

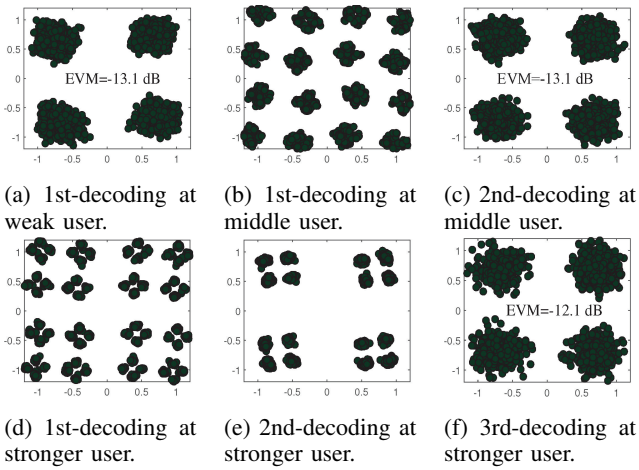
Case Study: Again, we use location 4 ($k = 4$) as an example to examine the proposed NOMA scheme. Fig. 11 presents the constellation of the decoded signals at the two users when NOMA and OMA are used, respectively. We have the following observations from the experimental data. (i) For the weak user, it has similar EVM in NOMA and OMA. Its extrapolated data rate is 3.0 Mbps in NOMA and 1.5 Mbps in OMA. This shows that NOMA has a significant throughput gain for the weak user. (ii) For the strong user, it achieves -20.0 dB EVM in NOMA and -24.9 dB EVM in OMA. Correspondingly, the extrapolated data rate for this user is 9.0 Mbps in NOMA and 6.0 Mbps in OMA. This shows that NOMA has a considerable throughput gain (50.0%) for the strong user as well. (iii) For the AP, the weighted sum rate is 27.0 Mbps in NOMA and 15.7 Mbps in OMA. This shows that our NOMA scheme has a 72.0% gain over OMA.

Results from All Locations: Fig. 12 presents the extrapolated data rate at each user when NOMA and OMA are used, respectively. The experimental results from the six locations corroborate our observations in the case study. On average, NOMA improves the data rate by 55.5% for the weak user, 40.7% for the strong user, 49.8% for the AP's weighted sum rate. Moreover, our proposed SIC outperforms ZF-SIC, yielding 35.7% data rate gain for the strong user.

E. Experimental Results of (2×3)*-NOMA*

Finally, we consider the case where the AP has two antennas and it serves three users. The weak user is placed at L_{k1}, the middle user is placed at L_{k2}, and the strong user is placed at L_{k3}, $k = 1, 2, \cdots, 6$.

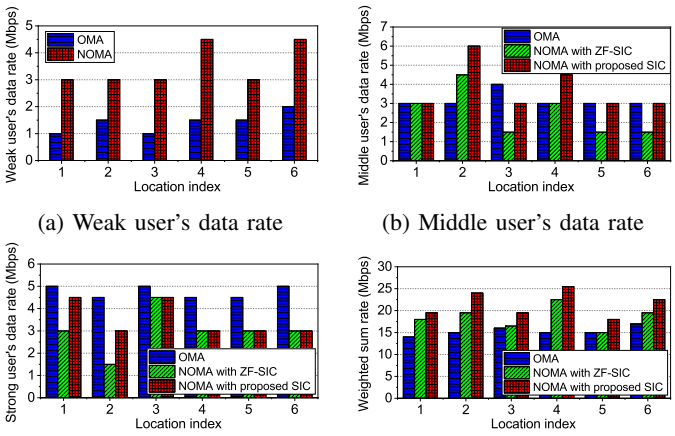
Case Study: Similar to the previous case studies, we place the users at location 4. Fig. 13 presents the constellation of the decoded signals at the users when NOMA is used. When OMA is used, the three stations achieve -14.5 dB, -21.4 dB, and -26.8 dB EVM, respectively. Based on the experimental results, we have following observations: (i) for the weak user, NOMA has a small EVM degradation (1.4 dB) compared to OMA. Its extrapolated data rate is 4.5 Mbps in NOMA and 1.5 Mbps in OMA. (ii) For the middle user, NOMA



(a) 1st-decoding at weak user. (b) 1st-decoding at middle user. (c) 2nd-decoding at middle user.

(d) 1st-decoding at stronger user. (e) 2nd-decoding at stronger user. (f) 3rd-decoding at stronger user.

Fig. 13: Constellation of the decoded signals in downlink NOMA transmission when the two-antenna AP serves three single-antenna users.



(a) Weak user's data rate (b) Middle user's data rate

(c) Strong user's data rate (d) AP's weighted sum rate

Fig. 14: Performance comparison of NOMA and OMA in downlink transmission when the two-antenna AP serves three single-antenna users.

has a considerable EVM degradation (8.3 dB) compared to OMA. Its extrapolated data rate is 4.5 Mbps in NOMA and 3.0 Mbps in OMA. (iii) For the strong user, NOMA has a significant EVM degradation (13.7 dB). Its extrapolated data rate is 3.0 Mbps in NOMA and 4.5 Mbps in OMA. (iv) For the AP, the weighted sum rate is 25.5 Mbps in NOMA and 15.0 Mbps in OMA. This means that NOMA has a 70.0% gain over OMA in terms of weighted sum rate.

Results from All Locations: Fig. 14 presents the extrapolated data rate at each user when NOMA and OMA are used. From the experimental results, we can see that NOMA significantly increases the weak user's data rate, slightly increases the middle user's data rate, and considerably decreases the strong user's data rate. On average over the six locations, NOMA increases the data rate by 147.1% for the weak user and by 18.4% for the middle user. However, it decreases the data rate of the strong user by 26.3%. For the AP, NOMA achieves a weighted sum rate of 21.5 Mbps and OMA achieves 15.3 Mbps, indicating a 40.5% improvement. Meanwhile, it is evident that our proposed SIC considerably outperforms ZF-SIC. It improves strong user's data rate by 16.7% and middle user's data rate by 50.0%.

F. Summary of Observations

Based on our experimental results, we have the following observations on NOMA: (i) NOMA can significantly increase the weak user's data rate when compared to OMA. This phenomenon has been observed in all the cases that we tested in our experiments. (ii) As expected, the use of NOMA will lead to a degradation for the strong user's data rate. But in overall, NOMA can greatly improve the weighted sum rate for the AP. (iii) Our proposed SIC method works in practice and it offers considerably better performance than ZF-SIC.

VII. Conclusions

In this paper, we proposed a NOMA scheme for WLANs and evaluated its performance in real-world wireless environments. Our NOMA scheme has three key components: precoder design, user grouping, and a new SIC method. We formulated the precoder design problem as an optimization problem and developed a minorization-majorization algorithm to pursue an efficient solution to it. Moreover, a robust SIC method has been proposed to decode the desired signal in the presence of strong interference. Our SIC method does not require channel estimation and is amenable to practical implementation. We have implemented the proposed NOMA scheme on a GNURadio-USRP2 testbed. Experimental results show that, compared with OMA, the proposed NOMA scheme can significantly improve the weak user's data rate and considerably improve the AP's weighted sum rate.

Appendix

Proof of Proposition 1

Consider the pairing strategies in Fig. 5. Suppose that the paired users are scheduled in the round-robin way over a set of time slots and that r_{bps} denotes the weighted sum rate yielded by the $(i, S/2+i)$ pairing strategy. Then, we have

$$r_{\text{bps}} = \sum_{i=1}^{S/2} \left[w_1 \log_2 \left(1 + \frac{(1-\alpha_i)\,|h_i|^2}{\alpha_i\,|h_i|^2 + \sigma^2} \right) + w_2 \log_2 \left(1 + \frac{\alpha_i\,|h_{S/2+i}|^2}{\sigma^2} \right) \right], \quad (22)$$

where $\alpha_i = p(h_i, h_{S/2+i})$, and $(1-\alpha_i)$ is then the normalized portion of AP's transmit power for the weak user, σ^2 is the normalized power of noise (w.r.t. the signal power), and w_1 and w_2 denote the weight assigned to the weak and strong users in a group, respectively.

To show that the $(i, S/2+i)$ pairing strategy yields the highest weighted sum rate among all possible pairing strategies, we argue that any permutation over this pairing strategy would lead to a decrease in the weighted sum rate. Without loss of generality, we assume that the permutation occurs for user pairs $(1, S/2+1)$ and $(2, S/2+2)$. After permutation,

the resulting pairs are $(1, S/2+2)$ and $(2, S/2+1)$. For the permuted user pairs, we let $\alpha_1' = p(h_1, h_{S/2+2})$ and $\alpha_2' = p(h_2, h_{S/2+1})$. Then, the change of weighted sum rate from the permutation can be written as follows:

$$\begin{aligned}
\Delta &= r_{bps} - r_{perm} \\
&= w_1 \log_2\left(1+\frac{(1-\alpha_1)\left|h_1\right|^2}{\alpha_1\left|h_1\right|^2+\sigma^2}\right)+w_2\log_2(1+\frac{\alpha_1}{\sigma^2}\left|h_{S/2+1}\right|^2) \\
&+ w_1 \log_2\left(1+\frac{(1-\alpha_2)\left|h_2\right|^2}{\alpha_2\left|h_2\right|^2+\sigma^2}\right)+w_2\log_2(1+\frac{\alpha_2}{\sigma^2}\left|h_{S/2+2}\right|^2) \\
&- w_1 \log_2\left(1+\frac{(1-\alpha_1')\left|h_1\right|^2}{\alpha_1'\left|h_1\right|^2+\sigma^2}\right)-w_2\log_2(1+\frac{\alpha_1'}{\sigma^2}\left|h_{S/2+2}\right|^2) \\
&- w_1 \log_2\left(1+\frac{(1-\alpha_2')\left|h_2\right|^2}{\alpha_2'\left|h_2\right|^2+\sigma^2}\right)-w_2\log_2(1+\frac{\alpha_2'}{\sigma^2}\left|h_{S/2+1}\right|^2),
\end{aligned} \tag{23}$$

where r_{perm} denotes the weighted sum rate after permutation.

Through algebraic operations, (23) can be rewritten as:

$$\begin{aligned}
\Delta r = w_1\log_2\left(\frac{\alpha_2'\left|h_2\right|^2+\sigma^2}{\alpha_2\left|h_2\right|^2+\sigma^2}\right)-w_1\log_2\left(\frac{\alpha_1'\left|h_1\right|^2+\sigma^2}{\alpha_1\left|h_1\right|^2+\sigma^2}\right)+ \\
w_2\log_2\left(\frac{\alpha_2\left|h_{S/2+2}\right|^2+\sigma^2}{\alpha_1'\left|h_{S/2+2}\right|^2+\sigma^2}\right)-w_2\log_2\left(\frac{\alpha_1\left|h_{S/2+1}\right|^2+\sigma^2}{\alpha_2'\left|h_{S/2+1}\right|^2+\sigma^2}\right).
\end{aligned} \tag{24}$$

Recall that the users are sorted in increasing order by their channel strength, i.e., $|h_1| \le |h_2| \le \left|h_{S/2+1}\right| \le \left|h_{S/2+2}\right|$. Since $p(h_w, h_s)$ is a non-increasing function with respect to $|h_s| / |h_w|$, we have $\alpha_1 \ge \alpha_1'$, $\alpha_2' \ge \alpha_2$, $\alpha_2' \ge \alpha_1$, and $\alpha_2 \ge \alpha_1'$. Then, the following two inequalities are imminent.

$$\frac{\alpha_2'\left|h_2\right|^2+\sigma^2}{\alpha_2\left|h_2\right|^2+\sigma^2} \ge \frac{\alpha_1'\left|h_1\right|^2+\sigma^2}{\alpha_1\left|h_1\right|^2+\sigma^2}, \tag{25}$$

$$\frac{\alpha_2\left|h_{S/2+2}\right|^2+\sigma^2}{\alpha_1'\left|h_{S/2+2}\right|^2+\sigma^2} \ge \frac{\alpha_1\left|h_{S/2+1}\right|^2+\sigma^2}{\alpha_2'\left|h_{S/2+1}\right|^2+\sigma^2}. \tag{26}$$

Based on (24), (25), and (26), it is evident that $\Delta r \ge 0$. This shows that any permutation on user pairing $(i, S/2+i)$ decreases the weighted sum rate. We therefore conclude that the $(i, S/2+i)$ pairing strategy yields the highest weighted sum rate. By the same token, we can prove that the $(i, S+1-i)$ pairing strategy yields the lowest weighted sum rate. We omit this part to converse space.

REFERENCES

[1] F. Fang, H. Zhang, J. Cheng, and V. C. Leung, "Energy-efficient resource allocation for downlink non-orthogonal multiple access network," *IEEE Trans. Commun.*, vol. 64, no. 9, pp. 3722–3732, 2016.

[2] Y. Fu, L. Salaün, C. W. Sung, and C. S. Chen, "Subcarrier and power allocation for the downlink of multicarrier NOMA systems," *IEEE Trans. Veh. Technol.*, vol. 67, no. 12, pp. 11833–11847, 2018.

[3] J. Zhao, Y. Liu, K. K. Chai, Y. Chen, and M. Elkashlan, "Joint subchannel and power allocation for NOMA enhanced D2D communications," *IEEE Trans. Commun.*, vol. 65, no. 11, pp. 5081–5094, 2017.

[4] M. F. Hanif, Z. Ding, T. Ratnarajah, and G. K. Karagiannidis, "A minorization-maximization method for optimizing sum rate in the downlink of non-orthogonal multiple access systems," *IEEE Trans. Signal Process.*, vol. 64, no. 1, pp. 76–88, 2016.

[5] N. Zhao, X. Pang, Z. Li, Y. Chen, F. Li, Z. Ding, and M.-S. Alouini, "Joint trajectory and precoding optimization for UAV-assisted NOMA networks," *IEEE Trans. Commun.*, vol. 67, no. 5, pp. 3723–3735, 2019.

[6] Z. Chen, Z. Ding, P. Xu, and X. Dai, "Optimal precoding for a QoS optimization problem in two-user MISO-NOMA downlink," *IEEE Commun. Lett.*, vol. 20, no. 6, pp. 1263–1266, 2016.

[7] H. Al-Obiedollah, K. Cumanan, J. Thiyagalingam, A. G. Burr, Z. Ding, and O. A. Dobre, "Energy efficient beamforming design for MISO non-orthogonal multiple access systems," *IEEE Trans. Commun.*, vol. 67, pp. 4117–4131, June 2019.

[8] Y. Liu, Z. Qin, M. Elkashlan, Y. Gao, and L. Hanzo, "Enhancing the physical layer security of non-orthogonal multiple access in large-scale networks," *IEEE Trans. Wireless Commun.*, vol. 16, no. 3, pp. 1656–1672, 2017.

[9] Y. Cao, N. Zhao, Y. Chen, M. Jin, L. Fan, Z. Ding, and F. R. Yu, "Privacy preservation via beamforming for NOMA," *IEEE Trans. Wireless Commun.*, vol. 18, pp. 3599–3612, July 2019.

[10] N. Zhao, D. Li, M. Liu, Y. Cao, Y. Chen, Z. Ding, and X. Wang, "Secure transmission via joint precoding optimization for downlink MISO NOMA," *IEEE Trans. Veh. Technol.*, vol. 68, pp. 7603–7615, Aug 2019.

[11] S. Verdu, *Multiuser Detection*. Cambridge, U.K.: Cambridge Univ. Press, 1998.

[12] 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, 2015.

[13] H. Xing, Y. Liu, A. Nallanathan, Z. Ding, and H. V. Poor, "Optimal throughput fairness tradeoffs for downlink non-orthogonal multiple access over fading channels," *IEEE Trans. Wireless Commun.*, vol. 17, no. 6, pp. 3556–3571, 2018.

[14] Z. Ding, Z. Yang, P. Fan, and H. V. Poor, "On the performance of non-orthogonal multiple access in 5G systems with randomly deployed users," *IEEE signal process. lett.*, vol. 21, no. 12, pp. 1501–1505, 2014.

[15] Z. Yang, Z. Ding, P. Fan, and N. Al-Dhahir, "A general power allocation scheme to guarantee quality of service in downlink and uplink NOMA systems," *IEEE Trans. Wireless Commun.*, vol. 15, no. 11, pp. 7244–7257, 2016.

[16] W. Yu, L. Musavian, and Q. Ni, "Link-layer capacity of NOMA under statistical delay QoS guarantees," *IEEE Trans. Commun.*, vol. 66, no. 10, pp. 4907–4922, 2018.

[17] J. A. Oviedo and H. R. Sadjadpour, "A fair power allocation approach to NOMA in multiuser SISO systems," *IEEE Trans. Veh. Technol.*, vol. 66, no. 9, pp. 7974–7985, 2017.

[18] X. Li, C. Li, and Y. Jin, "Dynamic resource allocation for transmit power minimization in OFDM-based NOMA systems," *IEEE Commun. Lett.*, vol. 20, no. 12, pp. 2558–2561, 2016.

[19] Y.-F. Liu, "Complexity analysis of joint subcarrier and power allocation for the cellular downlink OFDMA system.," *IEEE Wireless Commun. Lett.*, vol. 3, no. 6, pp. 661–664, 2014.

[20] N. Zhao, W. Wang, J. Wang, Y. Chen, Y. Lin, Z. Ding, and N. C. Beaulieu, "Joint beamforming and jamming optimization for secure transmission in MISO-NOMA networks," *IEEE Trans. Commun.*, vol. 67, no. 3, pp. 2294–2305, 2018.

[21] Z. Ding, P. Fan, and H. V. Poor, "Impact of user pairing on 5G nonorthogonal multiple-access downlink transmissions," *IEEE Trans. Veh. Technol.*, vol. 65, no. 8, pp. 6010–6023, 2015.

[22] F. Liu, P. Mähönen, and M. Petrova, "Proportional fairness-based power allocation and user set selection for downlink NOMA systems," in *Proc. IEEE Int. Conf. Commun. (ICC)*, pp. 1–6, 2016.

[23] X. Zhang, J. Wang, J. Wang, and J. Song, "A novel user pairing in downlink non-orthogonal multiple access," in *Proc. IEEE Int. Symp. Broadband Multimedia Syst. Broadcast.*, pp. 1–5, 2018.

[24] S. N. Datta and S. Kalyanasundaram, "Optimal power allocation and user selection in non-orthogonal multiple access systems," in *Proc. IEEE Wireless Commun. Netw. Conf. (WCNC)*, pp. 1–6, 2016.

[25] W. Liang, Z. Ding, Y. Li, and L. Song, "User pairing for downlink non-orthogonal multiple access networks using matching algorithm," *IEEE Trans. Commun.*, vol. 65, no. 12, pp. 5319–5332, 2017.

[26] Z. Ding, R. Schober, and H. V. Poor, "A general MIMO framework for NOMA downlink and uplink transmission based on signal alignment," *IEEE Trans. Wireless Commun.*, vol. 15, no. 6, pp. 4438–4454, 2016.

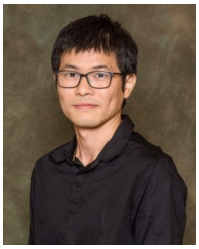
[27] A. Benjebbour, Y. Kishiyama, Y. Okumura, C.-H. Hwang, and I.-K. Fu, "Outdoor experimental trials of advanced downlink NOMA using smartphone-sized devices," in *Proc. IEEE VTC-Spring*, pp. 1–6, 2018.

[28] A. Benjebbour, K. Saito, A. Li, Y. Kishiyama, and T. Nakamura, "Non-orthogonal multiple access (NOMA): Concept, performance evaluation

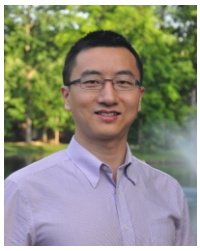
and experimental trials," in *Proc. Int. Conf. Wireless Netw. Mobile Commun. (WINCOM)*, pp. 1–6, 2015.
[29] A. Benjebbour, A. Li, K. Saito, Y. Saito, Y. Kishiyama, and T. Nakamura, "NOMA: From concept to standardization," in *Proc. IEEE Conf. Stand. Commun. Netw.*, pp. 18–23, 2015.
[30] A. Benjebbour and Y. Kishiyama, "Combination of NOMA and MIMO: Concept and experimental trials," in *Proc. Int. Conf. Telecommun.*, pp. 433–438, 2018.
[31] IEEE 802.11ac, "IEEE standard for information technology local and metropolitan area networks part 11: Wireless LAN medium access control (MAC) and physical layer (PHY) specifications amendment 5: Enhancements for higher throughput," *IEEE Std. 802.11ac*, 2014.
[32] S. Boyd and L. Vandenberghe, *Convex optimization*. Cambridge, U.K.: Cambridge Univ. Press, 2004.
[33] D. R. Hunter and K. Lange, "A tutorial on MM algorithms," *Amer. Statistician*, vol. 58, no. 1, pp. 30–37, 2004.
[34] F. Alizadeh and D. Goldfarb, "Second-order cone programming," *Math. program.*, vol. 95, no. 1, pp. 3–51, 2003.
[35] M. Andersen, J. Dahl, and L. Vandenberghe, "CVXOPT: Python software for convex optimization, version 1.1," 2015.
[36] Y. Nesterov and A. Nemirovskii, *Interior-Point Polynomial Algorithms in Convex Programming*. Philadelphia, PA, USA: SIAM, 1994.
[37] C. Shepard, H. Yu, N. Anand, E. Li, T. Marzetta, R. Yang, and L. Zhong, "Argos: Practical many-antenna base stations," in *Proc. ACM MobiCom*, pp. 53–64, 2012.

Pedram Kheirkhah Sangdeh received a B.Sc. degree in Electrical and Computer Engineering from Iran University of Science and Technology, Tehran, Iran, in 2011, and a M.Sc. degree in Electrical and Computer Engineering from the College of Engineering, University of Tehran, Tehran, Iran, in 2014. He is currently pursuing a Ph.D. degree with the Department of Electrical and Computer Engineering, Louisville, KY, USA. His research interests include performance analysis and implementation of innovative protocols for the next-generation of wireless networks.

Hossein Pirayesh received his B.Sc. degree in Electrical Engineering from Karaj Islamic Azad University, Karaj, Iran in 2013 and his M.Sc. degree in Electrical Engineering from Iran University of Science and Technology, Tehran, Iran in 2016. Since 2017, he has been working toward his Ph.D. degree in the Department of Electrical and Computer Engineering at the University of Louisville, Louisville, KY, USA. His current research is focused on wireless communications and networking, including theoretical analysis, algorithm and protocol design, and system implementation.

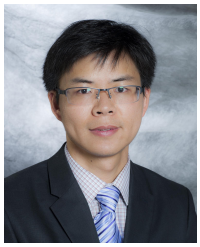


Qiben Yan (S'11-M'15) is an Assistant Professor in the Department of Computer Science and Engineering of Michigan State University. He was an Assistant Professor in the University of Nebraska-Lincoln between 2015 and 2019. He received his Ph.D. in Computer Science department from Virginia Tech, an M.S. and a B.S. degree in Electronic Engineering from Fudan University in Shanghai, China. His current research interests include wireless communication, wireless network security and privacy, mobile and IoT security.



Kai Zeng received the Ph.D. degree in electrical and computer engineering from the Worcester Poly-1239 technic Institute (WPI) in 2008. He was a Post-Doctoral Scholar with the Department of Computer Science, University of California at Davis (UCD), from 2008 to 2011. He was an Assistant Professor with the Department of Computer and Information Science, University of Michigan, Dearborn, from 2011 to 2014. He is currently an Associate Professor with the Department of Electrical and Computer Engineering, Cyber Security Engineering, and the Department of Computer Science, George Mason University. His current research interests include cyber-physical system security and privacy, 5G wireless network security, network forensics, and spectrum sharing networks. He was a recipient of the U.S. National Science Foundation Faculty Early Career Development (CAREER) Award in 2012. He received the Excellence in Post-Doctoral Research Award from UCD in 2011 and the Sigma Xi Outstanding Ph.D. Dissertation Award from WPI in 2008. He is an Associate Editor of the IEEE TRANSACTIONS ON INFORMATION FORENSICS AND SECURITY and the IEEE TRANSACTIONS ON COGNITIVE COMMUNICATIONS AND NETWORKING.

Wenjing Lou (F'15) is a Professor of Computer Science at Virginia Tech and a Fellow of the IEEE. She holds a Ph.D. in Electrical and Computer Engineering from the University of Florida. Her research interests cover many topics in the cybersecurity field, with her current research interest focusing on privacy protection techniques in networked information systems and cross-layer security enhancement in wireless networks. Prof. Lou is currently on the editorial boards of ACM/IEEE Transactions on Networking, IEEE Transactions on Mobile Computing, and Journal of Computer Security. She is the Steering Committee Chair of IEEE Conference on Communications and Network Security (IEEE CNS). She served as a program director at the US National Science Foundation (NSF) from 2014 to 2017.



Huacheng Zeng (M'15) is an Assistant Professor of Electrical and Computer Engineering at the University of Louisville, Louisville, KY. He holds a Ph.D. degree in Computer Engineering from Virginia Tech, Blacksburg, VA. He worked as Senior System Engineer at Marvell Semiconductor, Santa Clara, CA for two years. His research is on a broad topics in wireless networking and mobile computing, including algorithm and protocol design, interference management, physical-layer security, learning-based radar applications. He is a recipient of the NSF CAREER Award in 2019.