

# Throughput Analysis and User Barring Design for Uplink NOMA-Enabled Random Access

Wenjuan Yu, *Member, IEEE*, Chuan Heng Foh, *Senior Member, IEEE*, Atta ul Quddus, Yuanwei Liu, *Senior Member, IEEE*, Rahim Tafazolli, *Senior Member, IEEE*

**Abstract**—Being able to accommodate multiple simultaneous transmissions on a single channel, non-orthogonal multiple access (NOMA) appears as an attractive solution to support massive machine type communication (mMTC) that faces a massive number of devices competing to access the limited number of shared radio resources. In this paper, we first analytically study the throughput performance of NOMA-based random access (RA), namely NOMA-RA. We show that while increasing the number of power levels in NOMA-RA leads to a further gain in maximum throughput, the growth of throughput gain is slower than linear. This is due to the higher-power dominance characteristic in power-domain NOMA known in the literature. We explicitly quantify the throughput gain for the very first time in this paper. With our analytical model, we verify the performance advantage of NOMA-RA scheme by comparing with the baseline multi-channel slotted ALOHA (MS-ALOHA), with and without capture effect. Despite the higher-power dominance effect, the maximum throughput of NOMA-RA with four power levels achieves over three times that of the MS-ALOHA. However, our analytical results also reveal the sensitivity of load on the throughput of NOMA-RA. To cope with the potential bursty traffic in mMTC scenarios, we propose adaptive load regulation through a practical user barring algorithm. By estimating the current load based on the observable channel feedback, the algorithm adaptively controls user access to maintain the optimal loading of channels to achieve maximum throughput. When the proposed user barring algorithm is applied, simulations demonstrate that the instantaneous throughput of NOMA-RA always remains close to the maximum throughput confirming the effectiveness of our load regulation.

**Index Terms**—Massive machine type communication (mMTC), NOMA, random access, user barring.

This work was partly funded by grant N62909-17-1-2114 from the US Office of Naval Research Global. In addition, the authors would also like to acknowledge the support of the University of Surrey 5GIC & 6GIC (<http://www.surrey.ac.uk/ics>) members for this work.

W. Yu is with the School of Computing and Communications, InfoLab21, Lancaster University, Lancaster, U.K. (Email: [w.yu8@lancaster.ac.uk](mailto:w.yu8@lancaster.ac.uk)). C. Foh, A. Quddus and R. Tafazolli are with 5GIC & 6GIC, Institute for Communication Systems (ICS), University of Surrey, Guildford, U.K. (Emails: {[c.foh](mailto:c.foh@surrey.ac.uk), [a.quddus](mailto:a.quddus@surrey.ac.uk), [r.tafazolli](mailto:r.tafazolli@surrey.ac.uk)}@surrey.ac.uk). Y. Liu is with the School of Electronic Engineering and Computer Science, Queen Mary University of London, London, U.K. (Email: [yuanwei.liu@qmul.ac.uk](mailto:yuanwei.liu@qmul.ac.uk))

## I. INTRODUCTION

Massive machine type communication (mMTC) has been considered to be an imperative and challenging scenario in future communication networks, where millions of Internet of Things (IoT) devices per square kilometre are deployed to support the massive connectivity. It is predicted that machine-to-machine (M2M) connections will experience a noticeable growth, where the number of M2M connections will reach 3.9 billion by 2022 [1]. Sporadic access will be expected by a large quantity of devices to transmit small data payloads, which results in an unknown, random subset of active devices at a given transmission instance [2]. One of the major issues is how to utilize the limited radio resources to support a large number of devices transmitting small amounts of data, while maintaining low-latency access. Hence, 3GPP introduced Long Term Evolution (LTE) Category M1/M2 user equipment types and narrowband IoT (NB-IoT) to provide the massive connectivity for wide-area coverage, while non-3GPP technologies such as LoRa and Sigfox are also in operation as competing technologies [3].

We shall take the NB-IoT as an example. For the uplink transmissions, a contention-based random access (RA) procedure is performed for initial uplink grant which includes the four-step message handshake between IoT devices and evolved node base station (eNB). Firstly, in step (i), an IoT device starts its RA procedure by transmitting a preamble, i.e., Msg1, to the eNB on the Narrowband Physical RA CHannel (NPRACH). A preamble is composed of four symbol groups and each of them is transmitted on a different subcarrier determined by the fixed size frequency hopping [4]. If two or more IoT devices randomly choose the same initial subcarrier, the preamble sequence will collide but the eNB can still detect the preamble [5], [6]. Then, in step (ii), the devices that transmit the same preamble successfully detected by the eNB will receive the same RA response (RAR) message, i.e., Msg2, which contains the uplink resource grant and synchronization information. This causes the

message transmission collision in step (iii), because those devices that selected the same initial subcarrier in step (i) will transmit the radio resource control (RRC) connection requests, i.e., Msg3, on the same Narrowband Physical Uplink Shared CHannel (NPUSCH) resources [7]. Then, the eNB will fail to decode multiple RRC connection requests which results in backoff in the time domain.

In the mMTC scenario, preamble collision will be expected to occur more frequently compared to the traditional IoT scenarios due to the fact that a huge volume of low-cost, low-energy-consumption devices aim to access the limited amount of radio resources simultaneously. According to the RA procedure in NB-IoT network, it can be noticed that the increase of preamble collision may result in severe network congestion and long access delay. On the other hand, in recent years, non-orthogonal multiple access (NOMA) has shown to be a promising multiple access technique for future communication networks, due to its high spectral efficiency by allowing multiple users to simultaneously occupy the same radio resource, such as subcarrier/channel, code and time slot [8]–[10]. For power-domain NOMA, superposition coding is conducted at the transmitter side to perform user-multiplexing, while multiuser separation techniques such as successive interference cancellation (SIC) can be applied at the receiver side to decode the superimposed signals [11]. There are also spreading-based NOMA schemes which utilizes the user-specific low cross-correlation real/complex-valued spreading sequences to enable non-orthogonal transmissions [10]. To support massive connectivity, many existing papers aim to incorporate code-domain NOMA [12]–[14], or power-domain NOMA [15]–[21] into contention-based random access schemes such as pure ALOHA and slotted ALOHA. In this article, we adopt power-domain NOMA<sup>1</sup> as an enabling technology to enhance the conventional slotted-ALOHA-based RA procedure in NB-IoT and LTE-M, by allowing multiple IoT devices to transmit messages using the same uplink resource [22].

The power-domain NOMA-enabled random access, namely NOMA-RA, allows the users to randomly select one channel and one power level to transmit, by

defining a set of pre-determined power levels<sup>2</sup>. Then, the eNB with SIC technique will successively decode the received information, based on the received-power difference. Compared to the conventional RA schemes such as pure ALOHA and slotted ALOHA, this newly proposed NOMA-RA scheme is more spectrum-efficient. Focusing on the novel NOMA-RA scheme, some research papers have been published which evaluate the throughput performance, such as [15]–[21]. The combination of pure ALOHA and power-domain NOMA was considered in [15], which showed that the introduction of NOMA results in a higher throughput, compared to the conventional pure ALOHA. Since we focus on the multi-channel slotted ALOHA (MS-ALOHA) in this paper, the analytical results obtained in [15] cannot be simply applied in our scheme. In [16]–[19], throughput analysis was conducted, but the definition of throughput is different from ours, as well as the research methodology. For example, the average throughput in [17] was defined as the average achievable sum rate, given that all users randomly choose between two power levels for transmissions while the maximum number of users selecting each power level is constrained. In this article, we focus on the throughput that is defined as the average number of successfully transmitted packets from all users through all channels and all power levels, where the number of power levels can be any value, as well as the number of channels and users/devices. Similarly, in [20], it was assumed that there are in total two power levels in the non-orthogonal random access system, while our model, that considers any number of power levels, is more general. The throughput analysis given in [20] cannot be easily extended to a more general model with more than two power levels, due to the greatly increased computation complexity in throughput calculations. A simple and scalable analysis approach is absolutely necessary.

In [21], a user can decide its transmission power following a channel-inversion-based power control scheme to achieve a pre-determined received-power level where there are multiple power levels in total. The throughput was analyzed in [21], but only a lower bound was provided, instead of exact analytical expressions. Similarly, in [23], by incorporating power-domain NOMA into MS-ALOHA, the author proposed a lower bound for through-

<sup>2</sup>Here, the power level is the received-power level and we adopt the same power model given in [23]. In the literature, there are many research works focusing on the design of power control models, such as the channel-inversion-based ones [20], [21], [23], the power-backoff-based ones [18], [19], and others based on learning techniques [24], [25]. The research topic of designing power control model is beyond the scope of this paper, while different designs will not influence the conclusions and insights obtained in this paper.

<sup>1</sup>The power-domain NOMA is hereinafter referred to as NOMA for simplicity.

put defined as the average number of signals that can be successfully decoded, which is accurate only for some special cases. Considering a NOMA-RA scheme, an approximate throughput was given in [26] which takes into account packet collision as well as decoding errors due to low signal-to-interference-plus-noise ratio (SINR). Although the above-mentioned papers investigated NOMA-RA and evaluated its throughput performance, theoretical analysis on throughput that provides the exact analytical expressions is not given in the literature, as well as the analysis on optimal load. We believe that our throughput analysis, which gives the exact analytical expressions for both Binomial and Poisson arrivals for a NOMA-enabled random access scheme with any given number of power levels and radio resources, fills a research gap in the literature. Furthermore, we also quantifies the performance gain of NOMA-RA in terms of throughput, which can serve as a useful reference and practical guideline for designing favourable random access procedures in mMTC scenarios. The comparison between NOMA-RA and the conventional MS-ALOHA with capture effect is also provided in this paper, where with capture effect, the receiver has the ability to decode the strongest signal even if there are packet collisions on a given time-frequency resource [27].

In this article, we focus on the novel NOMA-RA scheme and theoretically study its performance gain in terms of throughput and optimal offered load for the very first time. It is well known that the maximum throughput of conventional slotted-ALOHA is approximately 0.368 achieved when the offered load<sup>3</sup> is 1, which indicates that successful transmissions happen only 36.8% of time in slotted-ALOHA [28]. Correspondingly, one interesting question that needs to be addressed is: what are the maximum throughput and optimal load for the NOMA-RA scheme? This paper addresses this problem and provides exact analytical expressions for throughput, as well as the analysis for optimal load that achieves maximum throughput. The comparison with the conventional MS-ALOHA (with and without capture effect) is also numerically shown. Our study indicates that the NOMA-RA scheme provides quite considerable performance gain, even compared to the MS-ALOHA with capture effect.

Apart from designing a more spectrum-efficient RA scheme, another challenging issue for mMTC lies in the traffic burstiness, which is critical when a large number of devices simultaneously activate. To deal with this problem, many techniques have been developed, such

<sup>3</sup>The offered load is defined as the expected number of packets attempted in a time slot.

as backoff-based mechanisms [29], access class barring (ACB) [30], and preamble barring [31]. Backoff mechanisms defer the re-transmissions of collided packets by a random time to prevent successive collisions. However, as the number of re-transmission attempts increases, the backoff delay may increase exponentially. The joint base station association and power control design was considered in [32], [33], which aims to maximize system-wide utility, minimize total transmit power, while achieving traffic load balancing. However, this approach requires cooperation among users and also deterministic resource assignments, which can be difficult for random access schemes like NOMA-RA. According to ACB in LTE-A [30], the eNB periodically broadcasts barring parameters including the barring rate (or called access probability in this paper) and the mean barring time. Then, at the beginning of RA procedure, each user determines its barring status with the information broadcast by the eNB, by drawing a uniform random number between 0 and 1 and comparing with the barring rate [34]. If this random number is less than or equal to the barring rate ( $0 \leq p \leq 1$ ), the user selects and transmits a preamble. Otherwise, the user waits for a random duration based on the mean barring time. In ACB, it is crucial to adaptively control the barring parameters. This is also an open issue for the novel NOMA-RA scheme. In this paper, we propose a user barring algorithm for NOMA-RA which avoid massive synchronized access demands by redistributing access requests of users through time. It can be applied at the eNB to adjust barring rate and enable adaptive congestion control based on the real-time traffic load observations. The main contributions of this article are summarized below.

- Focusing on the NOMA-RA scheme, we derive the exact analytical expressions for throughput by considering two different arrival models, i.e., Binomial and Poisson arrivals. Our throughput analysis reveals that an increase in the number of power levels leads to an increase in maximum throughput, but the rate of increase is slower than linear, due to the "higher-power dominance" effect in power-domain NOMA. Further, the optimal load which achieves the maximum throughput is also derived for Poisson arrivals.
- A comprehensive comparison between NOMA-RA and the conventional MS-ALOHA (with and without capture effect) is conducted, by utilizing the derived exact analytical expressions. Our presented analytical work quantifies the performance gain, defined as the ratio of maximum throughput achieved

by NOMA-RA to that of MS-ALOHA<sup>4</sup>. Interestingly, even with the "higher-power dominance" effect, NOMA-RA with four power levels can still achieve a maximum throughput that is three times higher than that of MS-ALOHA.

- To alleviate the traffic burstiness, a practical user barring algorithm is proposed for NOMA-RA where two steps are involved: (i) load estimation; (ii) adaptive user access control. The load estimation requires the observation of channel outcomes. We find that traffic load can be accurately estimated based on the instantaneous throughput and idle channels<sup>5</sup>. Precisely, the eNB observes the instantaneous throughput and the number of idle channels for a fixed barring period, which are then compared with the pre-calculated databases to estimate the current load. Based on the load estimation, the access probability can be adjusted for the next period to achieve optimal load. Simulations results, that are based on the practical 3GPP traffic model following Beta distribution [35], show that with our user barring algorithm applied, the instantaneous throughput always remains close to the maximum throughput.

The remainder of this paper is organised as follows. The system model is first introduced in Section II, which discusses the proposed NOMA-RA scheme, as well as the conventional MS-ALOHA. In Section III, we conduct the theoretical analysis on throughput and optimal load for NOMA-RA, as well as on the probability of channel being idle, which paves the way for the user barring access design. Based on the analytical results, the adaptive access control is studied and the pseudocode for a proposed user barring algorithm is provided in Section IV. Simulation results are included in Section V, followed by conclusions summarized in Section VI.

## II. SYSTEM MODEL

Let us consider the uplink transmissions for a system consisting of one eNB,  $N$  orthogonal channels<sup>6</sup> and  $U$  users. Time is discretized into slots of duration  $\tau$  corresponding to the packet transmission time. All devices are assumed to be synchronized to the slot boundary. At the beginning of each transmission slot, all users have

<sup>4</sup>Note that by default, MS-ALOHA refers to the scheme without the consideration of capture effect.

<sup>5</sup>We define an idle channel as one that at least one idle power level is observed during the SIC decoding.

<sup>6</sup>Here, one channel means one frequency resource, e.g., one sub-carrier in NB-IoT. Since this paper mainly focuses on the theoretical throughput based on probability analysis, the system model is more general, which is not limited to NB-IoT standards.

the same access probability  $P_{\text{access}}$  to transmit a single packet, where  $P_{\text{access}} \in [0, 1]$ . Then, within each slot, the transmission attempts can be modelled as  $U$  independent Bernoulli trials. The number of transmitted packets (or the number of active users) in one time slot,  $U_{\text{access}}$ , follows a Binomial distribution with the expectation  $E[U_{\text{access}}] = UP_{\text{access}}$ . For NOMA-RA scheme, each active user randomly selects one channel and one power level to transmit. There are  $L$  predetermined received-power levels for each channel, denoted by  $\gamma_1 > \gamma_2 > \dots > \gamma_L > 0$ , in an ordered manner. Assume that user  $k$  chooses the power level  $l$  and the channel  $n$  for random access. Denote by  $I_n$  the index set of active users transmitting through the channel  $n$ . Then, the received signal at the eNB over channel  $n$  can be written as

$$y_n = \sum_{k \in I_n} h_{n,k} \sqrt{\rho_{n,k}} s_{n,k} + n_n, \quad (1)$$

where  $h_{n,k}$ ,  $\rho_{n,k}$ , and  $s_{n,k}$  represents the channel gain, transmit power and signal from user  $k$  using channel  $n$ , respectively. Further,  $n_n \sim \mathcal{CN}(0, N_0)$  is the additive white Gaussian noise where  $N_0$  is the noise spectral density. Then, the noise power is  $\sigma^2 = N_0 B$  where  $B$  is the bandwidth per channel.

According to [23], after selecting the power level  $l$ , the user  $k$ 's transmission power is given as  $\rho_{n,k} = \gamma_l / |h_{n,k}|^2$  where the received-power level  $\gamma_l$  is given by  $\Gamma(V_l + \sigma^2)$ . Here,  $V_l = \sum_{m=l+1}^L \gamma_m$  and  $V_L = 0$ . Further,  $\Gamma$  is the target SINR when there exists only one packet at each power level. Then, after some algebraic manipulations, we can obtain that  $\gamma_l = \Gamma \sigma^2 (\Gamma + 1)^{L-l}$ ,  $\forall l = 1, \dots, L$ . Based on this power model, if only one active user exists at each power level, the SINR for the active user who choose  $\gamma_1$  becomes  $\frac{\gamma_1}{V_1 + \sigma^2}$ , which equals to  $\Gamma$ . Similarly, the SINR for the user choosing  $\gamma_2$  also equals to  $\Gamma$ . Hence, all the  $L$  signals can be decoded using SIC in the ascending order if the transmission rate is given by  $\log_2(1 + \Gamma)$ . To apply this power model, it is assumed that the channel gain between user  $k$  and eNB, i.e.,  $h_{n,k}$ , is perfectly known at user  $k$  as a priori knowledge, so that the transmission power is adjusted to guarantee that the received power level at the eNB is  $\gamma_l$ . In TDD system, the eNB first sends known pilot signals to the users on the forward link. Since channel reciprocity holds between the forward and reverse link in TDD system, each user can directly get an estimate of reverse link CSI. In FDD system, a pilot-based channel training scheme can be deployed to acquire an estimate of reverse link CSI at the eNB. Then, the eNB can feedback the obtained CSI estimation to the users [36].

Acquiring the channel state is a resource-consuming task in time-varying systems, especially when the number of users is massive, and the obtained information may be imperfect [37]. However, the imperfection in CSI estimate and the resource allocation between pilot transmission, CSI feedback and data transmission are beyond the scope of this study.

Note that other power control algorithms may also be applicable but are beyond the scope of this paper<sup>7</sup>. Due to the properties of NOMA, one can notice that the NOMA-RA scheme is more spectrum-efficient compared to the commonly utilized RA schemes because it allows multiple users to simultaneously transmit using the same time-frequency resource. As long as the packets occupying the same radio resource arrive at the eNB with different selected received-power levels, the eNB employed with SIC can successively decode all the received signals. In order to focus on the theoretical limit and isolate the performance impact of the SIC operation, we assume perfect SIC where the decoded signal can be perfectly removed without leaving residual interference. Note that the conventional RA schemes, such as pure ALOHA and slotted ALOHA, can only support at most one packet on one radio resource, which result in large collision probability and low throughput in heavy load scenarios. The NOMA-RA scheme can address this problem due to the introduction of a new dimension of power-domain. For a NOMA-RA system with  $L$  power levels in total, it can support at most  $L$  packets on one radio resource, which can effectively reduce the collision probability and result in higher throughput. However, since there is no central management and all active devices just randomly select one channel and one power level, there is still a probability of collision. Hence, it is important to obtain the exact analytical expressions for the achievable throughput for the NOMA-RA scheme.

Let us focus on the NOMA-RA scheme and take the user  $k$  as an example. Assume that for a given time slot, it chooses the channel  $n$  and the power level  $l$  to transmit. The successful packet transmission of user  $k$  occurs (or known as the case of "successful packet") when both the following conditions are satisfied:

- Only user  $k$ 's packet is transmitted on channel  $n$  choosing power level  $l$ .
- On channel  $n$ , all the signals which choose higher received-power levels, i.e., from level 1 to level  $l-1$ , can be successfully decoded.

<sup>7</sup>Since the main focus of this paper is not on signal processing, an unsatisfied power level due to the limitation of hardware ability is out of scope of this work. There are some power control algorithms designed in the literature like [23] which manage to reduce the transmission power range, but these are beyond the scope of this paper.

If more than two packets transmitted on the same channel arrive at the eNB with the same selected power level, this is called "power collision". All the packet decoding on this power level will fail which in turn fails the decoding on all the lower power levels. This is called the issue of "higher-power dominance", due to the fact that SIC is adopted at the receiver. Once the decoding for one power level is successful, the decoded signal is removed before decoding the next level [38]. But if the decoding on one power level fails, the signals on the lower power levels cannot be decoded due to high interference power. Hence, in order to successfully decode user  $k$ 's signal, all the signals on higher power levels need to be successfully decoded and removed, then the user  $k$ 's packet can be decoded by treating the remaining signals on lower power levels as interference. In this work, we consider that the received-power levels are pre-determined in an appropriate way which rules out the possibility of packets on lower power levels destroying the higher power level's decoding [23]. Hence, by assuming that user  $k$  transmits on the channel  $n$  choosing the power level  $l$ , only the following situations result in the user  $k$ 's transmission failure<sup>8</sup>: 1) On channel  $n$ , there is more than one packet choosing the power level  $l$ , where power collision happens; 2) Only the user  $k$  transmits on channel  $n$  using power level  $l$ , but some packets transmitted on this channel with higher power levels (from level 1 to level  $l-1$ ) fail the decoding.

#### A. Conventional MS-ALOHA without Capture Effect

For comparison purposes, here we also briefly introduce the conventional MS-ALOHA, without capture effect being considered. Focusing on the uplink transmissions for a system with one eNB,  $N$  orthogonal channels and  $U$  devices. Within each time slot, all  $U$  devices have the same access probability  $P_{\text{access}}$  to transmit, where the active devices will choose a channel at random [39]. We take the user  $k$  as an example. When capture effect is not considered, the successful packet transmission for user  $k$  in a time slot only occurs when it transmits a packet on an idle channel that there is no user transmitting. On the other hand, if two or more packets access the same channel in a given time slot, then there is a collision and the receiver obtains no information. According to [39], we note that when  $U$  packets are to access  $N$  channels, the average number of successfully transmitted packets is given by  $U \left(1 - \frac{1}{N}\right)^{U-1}$ . When there is a large number of users ( $U \rightarrow \infty$ ), Poisson distribution with a

<sup>8</sup>Note that in this paper, collisions are assumed to be the only source of transmission failure. The decoding errors due to channel outage will be taken into account in our future research.

parameter  $\lambda$  can be used to model the number of packets accessing each channel. Then, the throughput achieved for Poisson arrivals accessing  $N$  orthogonal channels is given as  $N\lambda e^{-\lambda}$ . By taking the first derivative of  $N\lambda e^{-\lambda}$  and setting it to 0, the optimal load  $\lambda^*$  can be obtained which equals to 1. This indicates that the best system performance is achieved when there is on average one packet attempting to access one time slot, for each channel.

### B. Conventional MS-ALOHA with Capture Effect

To provide a comprehensive study, here we also introduce and analyze the conventional MS-ALOHA with capture effect considered. For the conventional MS-ALOHA, the receiver has no SIC decoder, but may have the ability to decode the strongest signal even if there are packet collisions on one channel in a given time slot. This is referred to as the "capture effect". In this work, we adopt a perfect capture model, which means that if one of the  $U$  transmitters chooses a power level that is larger than that chosen by each of the remaining  $U - 1$  transmitters, then capture effect occurs and a packet is successfully received by the receiver [27]. In contrary, if two or more packets choose the highest power level among all packets, then there is no successful decoding. One can notice that even with capture effect considered, the conventional MS-ALOHA scheme still can only support at most one packet per channel within one given time slot.

Assuming Binomial arrival process on each channel, the throughput for MS-ALOHA with capture effect can be derived, which is then given by [40]

$$T_{cap} = N \sum_{U_i=1}^U \frac{\binom{U_i}{1} \sum_{g=0}^{L-2} (L-g-1)^{U_i-1}}{L^{U_i}} \binom{U}{U_i} \times \left(\frac{1}{N}\right)^{U_i} \left(1 - \frac{1}{N}\right)^{U-U_i}. \quad (2)$$

On the other hand, if Poisson arrival process with a parameter  $\lambda$  is assumed for each channel, the throughput for MS-ALOHA with capture effect is given by

$$T_{cap} = N \sum_{U_i=1}^{\infty} \frac{\binom{U_i}{1} \sum_{g=0}^{L-2} (L-g-1)^{U_i-1}}{L^{U_i}} \frac{\lambda^{U_i} e^{-\lambda}}{U_i!}. \quad (3)$$

In Section V, numerical results are provided to compare the throughput performance between the proposed NOMA-RA and the conventional MS-ALOHA (with and without capture effect).

## III. THEORETICAL ANALYSIS

Based on the above discussions, we now focus on the theoretical analysis and mathematical expressions for throughput and optimal load for the proposed NOMA-RA scheme. The probability of channel being idle is also investigated in this section. For simplicity, we consider that the access probability  $P_{\text{access}}$  is fixed and equals to 1, which means that all  $U$  devices will access the system and the number of active users  $U_{\text{access}} = U$ . In Section IV, we will consider an adjustable access probability and propose a user barring algorithm to further reduce the collision probability and maintain the maximum throughput.

### A. Throughput Analysis

Assume that there are  $U$  packets to independently and uniformly access  $N$  channels. For each channel, there are  $L$  power levels which can be chosen. Within a given time slot, each packet randomly selects one channel and one power level to access. We define throughput as the average number of successfully transmitted packets that are delivered from all users through all channels, denoted by  $T$ . Before deriving it, let us first analyze the conditional throughput for one channel, conditioned on the number of packets accessing this channel.

*Theorem 1:* Given that there are  $U_i$  packets sent to access channel  $i$ , the conditional throughput, defined as the average number of successfully transmitted packets on this channel, is given by  $E[S_i|U_i] = \sum_{S_i=1}^{\min(L, U_i)} S_i P(S_i|U_i)$ , where  $P(S_i|U_i)$  is the conditional probability of having  $S_i$  packets successful on channel  $i$ , shown in (4) at the top of the next page.

*Proof:* See Appendix A. ■

Here, condition  $S_i < \min(L, U_i)$  represents the cases where not all the  $U_i$  packets are successful, which implies that there are packet collisions. For these cases, the conditional probability  $P(S_i|U_i)$  can be derived by finding the probability that given  $U_i$  packets accessing the channel, there are  $S_i$  power levels occupied, each with a packet transmission producing  $S_i$  successful packets, with some idle<sup>9</sup> power levels appearing among the  $S_i$  successful packets, followed by a power collision occurring at the power level below all the  $S_i$  successful packets and those idle power levels if any, and finally all other possible transmissions below the power collision. Our approach to compute this probability involves in further conditioning the power level that the power

<sup>9</sup>Here, "idle" means there is no packet transmitted on this power level.

$$P(S_i|U_i) = \begin{cases} \sum_{g=0}^{L-1-S_i} \frac{\binom{U_i}{S_i} S_i! \binom{S_i+g}{S_i}}{L^{U_i}} \left( (L-S_i-g)^{U_i-S_i} - (L-S_i-g-1)^{U_i-S_i} \right) & \text{if } S_i < \min(L, U_i), \\ \binom{U_i}{S_i} \frac{S_i! \binom{L}{S_i}}{L^{U_i}} & \text{if } S_i = U_i \leq L, \\ 0 & \text{otherwise.} \end{cases} \quad (4)$$

collision appears. The full derivation of  $P(S_i|U_i)$  is given in Appendix A.

For the condition  $S_i = U_i \leq L$ , all  $U_i$  packets accessing channel  $i$  are successful. The conditional probability  $P(S_i|U_i)$  can be derived by finding the probability of distributing all  $U_i$  packets onto  $L$  power levels, where each level only serves at most one packet.

Theorem 1 focuses on the throughput of a single channel. We now extend Theorem 1 for the study of multiple orthogonal channels. Binomial distribution is normally used to model the number of successes in a fixed number of independent trials [28], [41]. With  $U$  packets randomly accessing  $N$  orthogonal channels in one time slot, the number of packets that each channel receives follows a Binomial distribution, where the probability that a channel is contended by  $U_i$  packets is simply  $P(U_i) = \binom{U}{U_i} \left(\frac{1}{N}\right)^{U_i} \left(1 - \frac{1}{N}\right)^{U-U_i}$ . Using Binomial distribution to describe the number of packets accessing a particular channel, we immediately obtain the following result.

*Theorem 2:* Consider that there are  $U$  packets randomly accessing  $N$  channels independently and unbiasedly. The throughput, defined as the average number of packets which are successfully transmitted on all channels in one time slot, is given by

$$T = N \sum_{U_i=1}^U E[S_i|U_i] \binom{U}{U_i} \left(\frac{1}{N}\right)^{U_i} \left(1 - \frac{1}{N}\right)^{U-U_i}, \quad (5)$$

where  $E[S_i|U_i]$  is the conditional throughput given in Theorem 1.

The throughput given in Theorem 2 considers a finite number of packets accessing  $N$  channels, where the number of packets appearing on a particular channel follows Binomial distribution. However, in mMTC scenarios, a massive number of low-cost, low-energy-consumption devices are expected (that is  $U \rightarrow \infty$ ). With a large number of potential arrivals onto a channel, Poisson arrivals are more appropriate. Hence, we extend our study to Poisson arrivals as follows.

*Theorem 3:* Assume that the number of packets

on each power level follows Poisson distribution with an arrival rate of  $\lambda$ . Then, the throughput, defined as the average number of packets which are successfully transmitted on all channels in one time slot, is given by

$$T = N \sum_{i=1}^L \lambda e^{-\lambda} (e^{-\lambda} + \lambda e^{-\lambda})^{i-1}. \quad (6)$$

*Proof:* See Appendix B. ■

Theorem 2 and Theorem 3 provide the exact analytical expressions for throughput for NOMA-RA scheme, by considering Binomial and Poisson arrivals, respectively. The accuracy of these expressions will be validated in Section V, as well as the comparison results with the conventional MS-ALOHA. We then focus on finding the optimal load for the NOMA-RA scheme.

### B. Optimal Load Analysis

As discussed earlier, Poisson distribution is more suitable for modelling packet arrivals from a large number of devices such as the mMTC scenarios. Our aim in this section is to find the optimal load which achieves the maximum throughput for NOMA-RA scheme. According to Theorem 3, we have that  $T = N \sum_{i=1}^L \lambda e^{-\lambda} (e^{-\lambda} + \lambda e^{-\lambda})^{i-1}$ . Optimal load is obtained when the first derivative of  $T$  with respect to  $\lambda$  equals to 0. Hence, by taking the first derivative, we get that

$$\begin{aligned} \frac{\partial T}{\partial \lambda} &= N \sum_{i=1}^L \frac{\partial \lambda e^{-\lambda} (e^{-\lambda} + \lambda e^{-\lambda})^{i-1}}{\partial \lambda} \\ &= N \sum_{i=1}^L \frac{\partial \lambda e^{-\lambda}}{\partial \lambda} (e^{-\lambda} + \lambda e^{-\lambda})^{i-1} \\ &\quad + \lambda e^{-\lambda} \frac{\partial (e^{-\lambda} + \lambda e^{-\lambda})^{i-1}}{\partial \lambda} \\ &= N \sum_{i=1}^L (e^{-\lambda} - \lambda e^{-\lambda}) (e^{-\lambda} + \lambda e^{-\lambda})^{i-1} \\ &\quad - (i-1) (\lambda e^{-\lambda})^2 (e^{-\lambda} + \lambda e^{-\lambda})^{i-2} \end{aligned}$$

$$= N \sum_{i=1}^L (e^{-2\lambda} - i\lambda^2 e^{-2\lambda}) (e^{-\lambda} + \lambda e^{-\lambda})^{i-2}. \quad (7)$$

By setting  $\frac{\partial T}{\partial \lambda}$  to zero, the optimal load  $\lambda^*$  can be obtained by numerically solving the following equation, by using *fzero* function in MATLAB or other numerical methods, e.g., bisection method.

$$N \sum_{i=1}^L \left( e^{-2\lambda^*} - i(\lambda^*)^2 e^{-2\lambda^*} \right) \left( e^{-\lambda^*} + \lambda^* e^{-\lambda^*} \right)^{i-2} = 0. \quad (8)$$

Note that the obtained  $\lambda^*$  is the optimal load for one power level, hence for a single channel with  $L$  power levels, the optimal load is simply  $\lambda^* L$ .

The determination of  $\lambda^*$  firstly permits us to compute the maximum throughput of NOMA-RA in order to compare with that of the baseline MS-ALOHA. Secondly, knowing how NOMA-RA should be optimally loaded for its best throughput performance, we can design a practical user barring algorithm which adaptively regulates the traffic load to maintain peak throughput performance. Before introducing the user barring algorithm, we first analyze the probability of the  $i^{\text{th}}$  channel being idle which will be utilized to estimate the actual load in the system.

### C. Probability of Idle Channel Analysis

Without loss of generality, we take the channel  $i$  as an example and define an indicator variable  $Z_i$ ,  $i \in \{1, 2, \dots, N\}$ , given by

$$Z_i = \begin{cases} 1 & \text{if any idle power level is observed} \\ & \text{during decoding on the } i^{\text{th}} \text{ channel,} \\ 0 & \text{otherwise.} \end{cases} \quad (9)$$

Here,  $Z_i = 1$ ,  $\forall i \in \{1, 2, \dots, N\}$ , indicates that there is at least one idle power level observed during the decoding<sup>10</sup> for the  $i^{\text{th}}$  channel. If  $Z_i = 1$ , we label the  $i^{\text{th}}$  channel as an "idle channel". Note that according to our definition, an "idle channel" does not necessarily mean that all power levels are idle on this channel. Instead, an "idle channel" indicates that at least one idle power level is observed before collision or before successfully decoding all power levels. The reason of defining "idle channel" in such a way is because when there is power collision, the successive decoding will terminate at the collision level. This means that the eNB can only label this channel relying on the information collected before

<sup>10</sup>Note that the decoding on one channel finishes when all power levels are successfully decoded or power collision happens.

the collision. Hence, we decide to investigate the probability of the  $i^{\text{th}}$  channel being idle, i.e.,  $P(Z_i = 1)$ , denoted by  $P_{\text{idle}}$ , which is later utilized to distinguish between light load and heavy load in the system. In the following theorem, we provide the analytical expression for the probability of channel being idle, i.e.,  $P_{\text{idle}}$ .

*Theorem 4:* Assume that the number of packets on each power level follows Poisson distribution with parameter  $\lambda$ . The probability of the channel being idle is given by

$$P_{\text{idle}} = (e^{-\lambda} + \lambda e^{-\lambda})^L - (\lambda e^{-\lambda})^L + (1 - e^{-\lambda} - \lambda e^{-\lambda}) \times \left( \sum_{i=2}^L (e^{-\lambda} + \lambda e^{-\lambda})^{i-1} - (\lambda e^{-\lambda})^{i-1} \right). \quad (10)$$

*Proof:* See Appendix C. ■

Theorem 4 can be applied directly to the case of multiple channels if users independently and unbiasedly choose a particular channel to access at a given time. Then, in the following section, we shall introduce the proposed user barring algorithm.

## IV. USER BARRING ALGORITHM

For mMTC scenarios, traffic overloading on the RA channel is a challenging issue, which can be especially critical when a large number of devices simultaneously activate, e.g., sensors reconnecting after a power outage [42]. This simultaneous triggering causes bursty arrivals, which can significantly degrade the performance of RA protocols. To alleviate the traffic burstiness, many techniques have been developed, such as the ACB proposed by 3GPP [30] and preamble barring [31]. In previous sections, the access probability, i.e.,  $P_{\text{access}}$ , was assumed to be fixed and equals to 1 for all users. In this section, we aim to propose a practical user barring algorithm to perform access control by adaptively adjusting the access probability  $P_{\text{access}}$  for all devices.

To be more practical, we assume that the access probability  $P_{\text{access}}$  remains fixed within  $\Lambda$  time slots and will be updated for the next period of  $\Lambda$ . Once the access probability  $P_{\text{access}}$  is broadcast to all users, each user decides whether to participate in the channel access for the period immediately after the broadcast with probability  $P_{\text{access}}$ , or remains silence otherwise. Then, the transmission attempts can be modelled as multiple independent Bernoulli trials, indicating that the number of active users for the next period of  $\Lambda$ , i.e.,  $U_{\text{access}}$ , follows a Binomial distribution with the expectation  $E[U_{\text{access}}] = \Lambda P_{\text{access}}$ . Note that  $P_{\text{access}}$ ,  $U_{\text{access}}$ , and  $\mathcal{A}_{\text{access}}$  will remain fixed for the next period of  $\Lambda$ , where  $\mathcal{A}_{\text{access}}$  is the index set of active users that decided to participate in channel access, i.e.,  $|\mathcal{A}_{\text{access}}| = U_{\text{access}}$ .



Then, within each time slot for the period of  $\Lambda$ , all  $U_{\text{access}}$  active users will randomly access  $N$  channels and  $L$  power levels, following the NOMA-RA scheme. Note that without central management and coordination, the eNB has no prior knowledge of the number of users accessing the system, i.e.,  $U_{\text{access}}$ . It can only guess the current load after receiving and decoding all signals on the uplink RA channel. Hence, the current load needs to be estimated first at the eNB, followed by the adaptive access control for the next period, with the aim of approaching optimal load.

#### A. Load Estimation

One of the key designs of user barring algorithm is load estimation. Being able to accurately estimate the load permits opportunity to regulate the loading of channel. In our analytical study, we have derived the optimal load for NOMA-RA to achieve peak throughput performance. According to the throughput results of NOMA-RA (see also Fig. 1), it can be noticed that observing instantaneous throughput alone is insufficient to estimate the load as a lightly and heavily loaded channel can produce the same throughput level causing ambiguity. Fortunately, we find that the appearance of idle power levels can be used to indicate light and heavy traffic conditions, and thus resolving the ambiguity. Besides, the observation of instantaneous throughput and the appearance of idle power levels does not incur additional signaling.

In our design, when the eNB receives and decodes all signals, it counts the number of successful packets over all channels and all power levels. After a period of  $\Lambda$ , the normalized instantaneous throughput  $T_{\text{insta}}$  can be calculated by  $\sum_{t=1}^{\Lambda} \sum_{i=1}^N S_{\text{insta}}^{i,t} / N / \Lambda$ , where  $S_{\text{insta}}^{i,t}$  is the number of successful packets on channel  $i$  in time slot  $t$ . According to Section III-A, we have obtained the analytical expressions for throughput, which provides the average number of packets that can be successfully transmitted. By comparing  $T_{\text{insta}}$  with the pre-calculated throughput matrix  $\mathcal{T}$ , which stores the discrete throughput values for various discretized load values, we can find two possible estimates for the number of active users, denoted by  $\tilde{U}_{\text{access}}^1$  (in light load region) and  $\tilde{U}_{\text{access}}^2$  (in heavy load region).

Then, one interesting phenomenon that has emerged is that when the offered load becomes heavier, i.e., when there are more packets expected on one channel, the number of power levels that remain unoccupied becomes less. This indicates that the probability of the channel being idle, i.e.,  $P_{\text{idle}}$ , analyzed in Theorem 4, is a monotonically decreasing function with the offered

load. Hence, due to the monotonic property of  $P_{\text{idle}}$ , a unique "load threshold"  $P_{\text{idle}}^{\tau}$  can be obtained which is defined as the probability of idle channel achieved at the optimal load, i.e.,  $P_{\text{idle}}^{\tau} = P_{\text{idle}}|_{\lambda=\lambda^*}$ . Apparently, the load threshold  $P_{\text{idle}}^{\tau}$  can be used to distinguish between light load and heavy load. Then, when the eNB decodes all the received signals during the period  $\Lambda$ , it will also observe the relative frequency of idle channels, i.e.,  $P_{\text{idle}}^{\text{insta}}$ , given by  $\sum_{t=1}^{\Lambda} \sum_{i=1}^N \mathbb{1}_{t,i}(\text{idle}) / N / \Lambda$ , where  $\mathbb{1}_{t,i}(\cdot)$  is an indicator function for channel  $i$  in time slot  $t$ . After a period of  $\Lambda$ , the obtained  $P_{\text{idle}}^{\text{insta}}$  will be compared with the pre-calculated threshold  $P_{\text{idle}}^{\tau}$ . If the observed  $P_{\text{idle}}^{\text{insta}}$  is larger than the threshold  $P_{\text{idle}}^{\tau}$ , it reveals that the system has more idle channels with a higher probability. In this case, we consider that the current load is light and the load estimate is  $\tilde{U}_{\text{access}}^1$ . Otherwise, we consider it to be in heavy load and the load estimate is  $\tilde{U}_{\text{access}}^2$ .

#### B. User Barring

Note that in order to conduct adaptive access control for the NOMA-RA system, there are mainly two steps involved: a) the eNB collects required information to calculate  $T_{\text{insta}}$  and  $P_{\text{idle}}^{\text{insta}}$  for the last period of  $\Lambda$ , which is then compared with  $\mathcal{T}$  and  $P_{\text{idle}}^{\tau}$  to estimate the current load; b) based on the estimated load, the eNB adjusts the access probability  $P_{\text{access}}$  for the next period of  $\Lambda$ , with the aim of maintaining the optimal load. In the above analysis, we have introduced how to utilize the collected information at the eNB to perform load estimation. In the following, we will focus on the second step and investigate how to adjust  $P_{\text{access}}$  for the next period.

Recall that if the instantaneously obtained  $P_{\text{idle}}^{\text{insta}}$  is larger than  $P_{\text{idle}}^{\tau}$ , it indicates that compared to the optimal load scenario, the current system has a higher probability to have more idle channels. In this case, we consider it to be in light load and the access probability can be increased for the next period, given by  $P_{\text{access}} = \min(1, P_{\text{access}} \frac{U^*}{\tilde{U}_{\text{access}}^1})$ , where  $U^*$  is the optimal number of active users (or called optimal load). In contrary, if  $P_{\text{idle}}^{\text{insta}}$  is smaller than  $P_{\text{idle}}^{\tau}$ , we consider the system to be in heavy load and the access probability needs to be reduced for the next period, given by  $P_{\text{access}} = \min(1, P_{\text{access}} \frac{U^*}{\tilde{U}_{\text{access}}^2})$ . The pseudocode for the complete user barring algorithm is given in Algorithm 1.

Let us analyze the computational complexity of the proposed Algorithm 1. Assume that the complexity of SIC at each receiver is a constant  $S$ . Since there are two nested loops from 1 to  $\Lambda$  and from 1 to  $N$  respectively, their complexity will be  $\mathcal{O}(\Lambda N)$  or simply  $\mathcal{O}(\Lambda N)$ . Then, line 17 can be implemented using two

binary search algorithms (one for the light load region and the other for the heavy load region) that have a complexity of  $2\mathcal{O}(\log(\mathcal{T}_{\text{size}}))$ . Here,  $\mathcal{T}_{\text{size}}$  is the size of the pre-calculated throughput matrix  $\mathcal{T}$  that depends on the discretization of offered load. For example, in our simulations, the load region is assumed to be  $[0, 20]$  and the discretized load interval is 0.2, which indicates that there are in total 101 discrete values, i.e.,  $\mathcal{T}_{\text{size}} = 101$ . Hence, one can notice that the complexity of line 17 is a constant which does not depend on the number of users, channels, or time slots. Therefore, the computational complexity of Algorithm 1 is  $\mathcal{O}(\Lambda N)$ . It can be noticed that the complexity of our proposed algorithm does not rely on the number of users in the system or the actual load in this instant. In other words, even if a massive number of devices become activated in an instant, the time complexity of our user barring algorithm does not proportionally increase.

## V. NUMERICAL RESULTS

In this section, numerical results are provided to validate the accuracy of the proposed analytical expressions in previous sections and further investigate the proposed NOMA-RA scheme. The performance of the proposed user barring algorithm will also be discussed through simulations. Firstly, theoretical analysis for throughput and optimal load of the proposed NOMA-RA scheme will be validated by comparing with Monte Carlo results. To get Monte Carlo results, we let  $U$  packets randomly choose power levels and channels. Then, for one experiment, the number of successfully transmitted packets is calculated, while the throughput can be found after conducting a large number of realizations. The analytical expressions for both arrival models, i.e., Binomial model and Poisson process, will be confirmed. Furthermore, the performance gain of NOMA-RA is also explicitly shown in this section, compared to the conventional MS-ALOHA (with and without capture effect). In the following, we start from the throughput performance for NOMA-RA, in comparison with the conventional MS-ALOHA without capture effect.

### A. Throughput Performance

Fig. 1 shows the normalized throughput, defined as  $T/N$ , versus the offered load ( $\lambda L$  or  $U/N$ ), for the proposed NOMA-RA scheme with different number of power levels. Here, the offered load is defined as the average number of packets per channel. For Binomial arrivals, the offered load is  $U/N$ , while for Poisson arrivals, the offered load per channel is  $\lambda L$ . The normalized throughput for Poisson arrivals are shown in solid

---

### Algorithm 1 User Barring Algorithm

---

**Input:**

- $N, L, U, \Lambda, P_{\text{access}}, \mathcal{A}_{\text{access}}, \mathcal{T}, P_{\text{idle}}^{\tau}, U^*$ .
- 1: **for**  $t \leftarrow 1$  to  $\Lambda$  **do**
  - 2: All  $U$  users generate random probabilities and compare with  $P_{\text{access}}$  to form  $\mathcal{A}_{\text{access}}$ ;
  - 3: Each active user in  $\mathcal{A}_{\text{access}}$  randomly chooses a channel and a power level to access;
  - 4: **for**  $i \leftarrow 1$  to  $N$  **do**
  - 5: Perform SIC at the receiver;
  - 6: **if** there is no power collision **then**
  - 7: Count the number of successful decoding, i.e.,  $S_{\text{insta}}^{i,t}$ ;
  - 8: Label it as an "idle channel" if any idle power level is noticed;
  - 9: **else**
  - 10: Count the number of successful decoding, i.e.,  $S_{\text{insta}}^{i,t}$ ;
  - 11: Label it as an "idle channel" if any idle level is noticed before collision;
  - 12: **end if**
  - 13: **end for**
  - 14: **end for**
  - 15: Calculate the normalized instantaneous throughput  $T_{\text{insta}}$ , i.e.,  $\sum_{t=1}^{\Lambda} \sum_{i=1}^N S_{\text{insta}}^{i,t} / N / \Lambda$ ;
  - 16: Calculate the relative frequency  $P_{\text{idle}}^{\text{insta}}$ , i.e.,  $\sum_{t=1}^{\Lambda} \sum_{i=1}^N \mathbb{1}_{t,i}(\text{idle}) / N / \Lambda$ ;
  - 17: Compare the actual throughput  $T_{\text{insta}}$  with  $\mathcal{T}$  to find  $\tilde{U}_{\text{access}}^1$  and  $\tilde{U}_{\text{access}}^2$ ;
  - 18: **if**  $P_{\text{idle}}^{\text{insta}} \geq P_{\text{idle}}^{\tau}$  **then**
  - 19:  $P_{\text{access}} = \min(1, P_{\text{access}} \frac{U^*}{\tilde{U}_{\text{access}}^1})$ ; % guess it is in light load
  - 20: **else**
  - 21:  $P_{\text{access}} = \min(1, P_{\text{access}} \frac{U^*}{\tilde{U}_{\text{access}}^2})$ ; % guess it is in heavy load
  - 22: **end if**
- Output:**  $P_{\text{access}}$
- 

lines, calculated using the analytical expressions given in Theorem 3. The throughput curves for Binomial arrivals are given using dots, calculated using the expressions given in Theorem 2, while Monte Carlo results are shown in dashed lines. To plot this figure, it is assumed that  $N = 10$ . Fig. 1 first confirms the accuracy of analytical expressions derived for the throughput for the NOMA-RA scheme, given in Theorem 2 and Theorem 3. Further, from Fig. 1, we can note that similar to the conventional MS-ALOHA, the normalized throughput

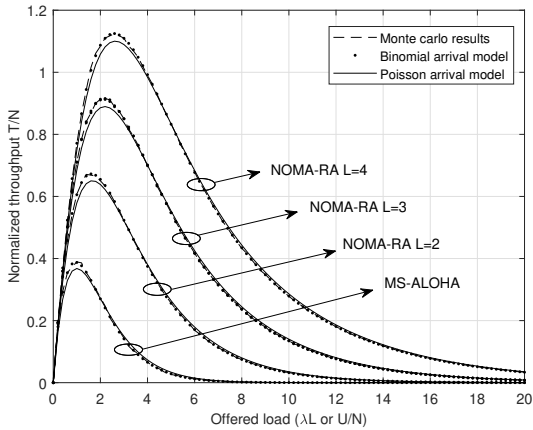


Fig. 1: Normalized throughput versus offered load for NOMA-RA and MS-ALOHA.

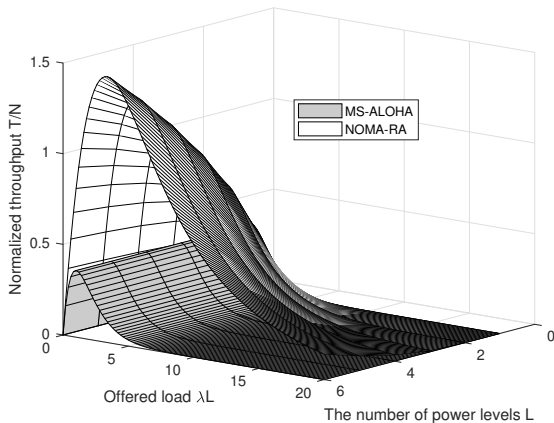


Fig. 2: Normalized throughput versus  $L$  and  $\lambda L$  for NOMA-RA and MS-ALOHA.

for the proposed NOMA-RA first increases with the offered load and then gradually decreases after reaching its peak value. This confirms that there is one unique maximum throughput and optimal load. When the number of power levels  $L = 1$ , the proposed NOMA-RA scheme is exactly the same with the conventional MS-ALOHA, offering a normalized maximum throughput of 0.368 achieved at the optimal load being 1. When there are more power levels, i.e.,  $L > 1$ , NOMA-RA achieves better throughput performance than MS-ALOHA.

To provide more comprehensive studies, Fig. 2 includes the plots for the normalized throughput versus the number of power levels  $L$  and the offered load  $\lambda$ , for the proposed NOMA-RA and the conventional MS-ALOHA. From this figure, we can note that the novel

NOMA-RA scheme always achieves higher normalized throughput than the conventional MS-ALOHA, regardless of the given offered load. On the other hand, it further shows that the normalized throughput for NOMA-RA gradually increases with the number of power levels  $L$ , while that of the conventional MS-ALOHA remains the same. Note that the performance metric considered in this figure is the normalized throughput, defined as  $T/N$ , which indicates that the impact of multiple channels has been eliminated. This is due to the reason that the linear increase in the number of independent channels, i.e.,  $N$ , results in a linear increase in the throughput (un-normalized). Hence, given a fixed load, the normalized throughput for the novel NOMA-RA scheme gradually increases with the number of power levels, while the normalized throughput remains the same for the conventional MS-ALOHA since its performance increase only relies on the number of channels.

It is noted that the introduction of multiple power levels in NOMA-RA brings another dimension. From the discussions and figures provided above, we can notice that although packet collisions still happen, the proposed NOMA-RA scheme achieves much better throughput performance, compared to the MS-ALOHA (without capture effect). Then, one intriguing question is: how many more channels would be required for MS-ALOHA to achieve the same performance as the novel NOMA-RA? To answer this question, we plot Fig. 3 which includes the curves of throughput (unnormalized), i.e.,  $T$ , versus the offered load, for both schemes. Specifically, it is assumed that there are two pre-determined power levels and four channels available for NOMA-RA. Then, we aim to find the required number of channels for MS-ALOHA to achieve the same throughput performance. Fig. 3 shows that at least eight channels are required for the conventional MS-ALOHA, in order to achieve a similar maximum throughput as the given NOMA-RA scheme. This confirms that the proposed NOMA-RA is much more spectrum-efficient as it only needs half of the radio resources to achieve a similar throughput performance.

### B. The Impact of Capture Effect in MS-ALOHA

The conventional MS-ALOHA discussed above will fail the packet decoding if packet collision happens on one channel, which is a scheme without the consideration of capture effect. According to the perfect capture model discussed in Section II-B, if there is one received packet which has a higher power than the others, capture effect occurs and this packet can be successfully decoded. By utilizing the analytical results given in Section II-B, it

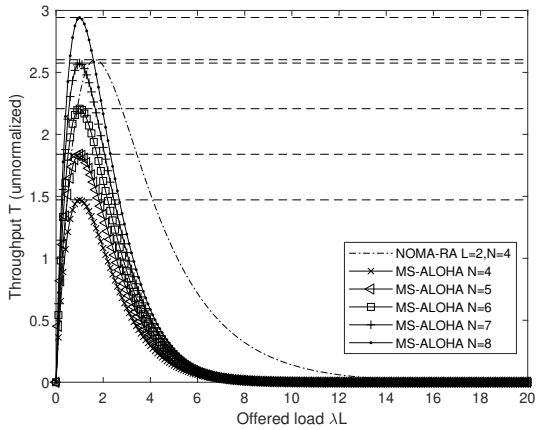


Fig. 3: Throughput (unnormalized) versus offered load  $\lambda L$  for NOMA-RA and MS-ALOHA.

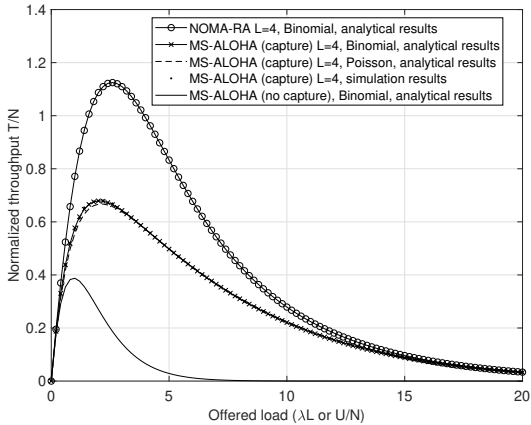


Fig. 4: Normalized throughput versus offered load for NOMA-RA and MS-ALOHA (with and without capture effect).

is interesting to check if the novel NOMA-RA is still superior and how much performance gain we can get, compared to the conventional MS-ALOHA with capture effect. Hence, we plot Fig. 4 which aims to compare the normalized throughput for three schemes, i.e., the novel NOMA-RA, the MS-ALOHA with capture effect and the MS-ALOHA without capture effect. From this figure, one can first notice that the proposed NOMA-RA still achieves better performance and the performance gain is quite considerable, even compared to the MS-ALOHA with capture effect. This is mainly because even with capture effect, the conventional MS-ALOHA still can only support at most one packet per channel within one time slot, while the novel NOMA-RA has

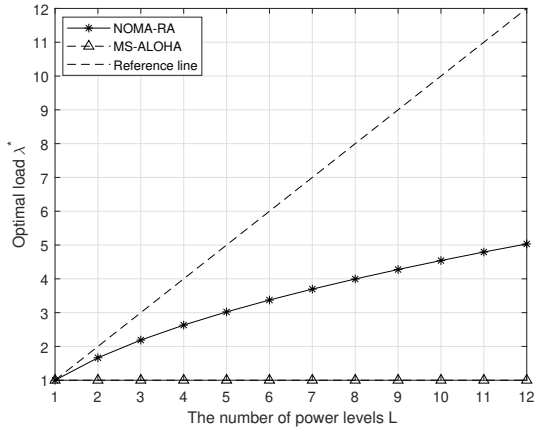
the possibility to support  $L$  packets per channel in one time slot. Hence, the throughput achieved by NOMA-RA can be larger. Fig. 4 also confirms the accuracy of the analytical expressions for the MS-ALOHA with capture effect given in Section II-B, which match with the Monte Carlo simulations.

### C. NOMA-RA Performance Gain

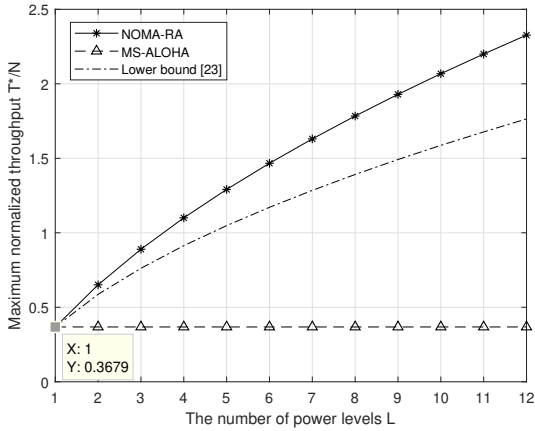
In order to investigate the optimal load and the maximum throughput for the proposed NOMA-RA, we include Fig. 5(a) and Fig. 5(b) that are plotted using the analytical results given in Sections III-A and III-B. The comparison results between NOMA-RA and MS-ALOHA are also given. Fig. 5(a) shows the optimal load  $\lambda^*$  versus the number of power levels  $L$  for both schemes<sup>11</sup>. From Fig. 5(a), we can first notice that the optimal load for the NOMA-RA scheme starts at the value of 1 and gradually increases with the number of power levels, while for the conventional MS-ALOHA, the optimal load is always achieved at 1. This indicates that the proposed NOMA-RA is more superior than MS-ALOHA since it has the ability to support a larger range of offered load, as we have noticed in Fig. 1. Interestingly, the optimal load does not increase linearly as the number of power level increases. This is due to the "higher-power dominance" effect in power domain NOMA-RA, resulting from the successive decoding technique.

Fig. 5(b) plots the maximum normalized throughput  $T^*/N$  versus the number of power levels  $L$ , for NOMA-RA, MS-ALOHA, and a lower bound on NOMA-RA throughput given in [23]. As we mentioned above, when  $L = 1$ , NOMA-RA is exactly the same with the conventional MS-ALOHA (without capture effect), which can also be confirmed in Fig. 5(b). Furthermore, similar to Fig. 5(a), Fig. 5(b) shows that the maximum normalized throughput for NOMA-RA increases with  $L$ , while that of MS-ALOHA remains the same. In [23], by considering a Poisson arrival with a parameter  $\lambda$ , a lower bound on the NOMA-RA throughput is given. By iterating through all feasible values of  $\lambda$ , a lower bound on the maximum throughput for NOMA-RA can be found, which is plotted in Fig. 5(b) and shown in dash-dotted line. We can clearly notice that the analytical expression given in [23] only serves as a lower bound and cannot reflect the exact throughput performance of NOMA-RA. Also, the lower bound given in [23] cannot clearly show the exact performance gain of NOMA-RA, compared to the conventional RA schemes. This

<sup>11</sup>The number of power levels is up to 12 in Fig. 5(a), Fig. 5(b) and Fig. 6. These figures only serve as theoretical guidance.



(a) Optimal load



(b) Maximum normalized throughput

Fig. 5: Optimal load and maximum normalized throughput versus  $L$ .

demonstrates the importance of the analysis of exact throughput expressions and the calculation of optimal load given in our paper.

In order to clearly show the performance gain of NOMA-RA, we plot Fig. 6 which shows its performance gain versus the number of power levels. Here, the performance gain is defined as the ratio of maximum throughput achieved by NOMA-RA to maximum throughput supported by MS-ALOHA. Fig. 6 first shows that the performance gain monotonically increases with the number of power levels. To clearly show its increase rate, we provide a reference line plotting the function  $y = x$ , where  $x \in [1, 12]$ . Fig. 6 shows that the maximum throughput achieved by NOMA-RA triples the conventional MS-ALOHA, when there are four power

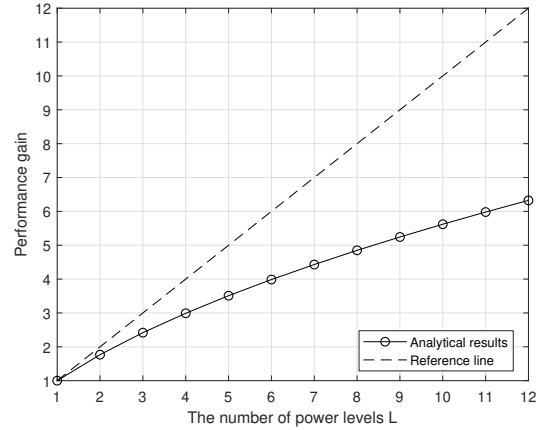


Fig. 6: The performance gain of NOMA-RA versus  $L$ .

levels. If we further increase the number of power levels, e.g.,  $L = 6$ , the maximum throughput achieved by NOMA-RA is four times larger than that of MS-ALOHA, which can be considered to be a substantial improvement. Furthermore, Fig. 5(a) and Fig. 6 also show that when the number of power levels increases, there is a widening gap between NOMA-RA and the reference line. This is because the growth of throughput gain is slower than linear which also becomes slower and slower, while the reference line linearly increases. Therefore, the gap gradually widens.

#### D. User Barring Performance - Ideal Environment

Note that we have investigated the proposed NOMA-RA and validated the accuracy of theoretical analysis given in previous sections. In this subsection, we consider that the access probability is adaptive and study the performance of user barring algorithm given in Section IV. More specifically, in this section, an ideal environment is considered where the impact of channel fading and decoding errors is excluded to specifically focus on the impact of transmission collisions. In Section V-E, we will investigate the performance of the proposed user barring algorithm in a more practical scenario, based on 3GPP macro propagation model [43].

Recall that in order to distinguish between light load and heavy load, the probability of idle channel have been introduced. Fig. 7 is plotted to validate the accuracy of analytical expressions for  $P_{\text{idle}}$  given in Theorem 4. To plot this figure, it is assumed that  $N = 10$  and  $L = 4$ . From this figure, it can be first noticed that the analytical results perfectly match with Monte Carlo simulations, for both probabilities. Further, it shows that the probability of channel being idle starts at the maximum value

TABLE I: Total number of users in the system

Time	$[0, 40\Lambda - 1]$	$[40\Lambda, 80\Lambda - 1]$	$[80\Lambda, 120\Lambda - 1]$	$[120\Lambda, 160\Lambda - 1]$	$[160\Lambda, 200\Lambda - 1]$
$U$	15	83	156	199	200

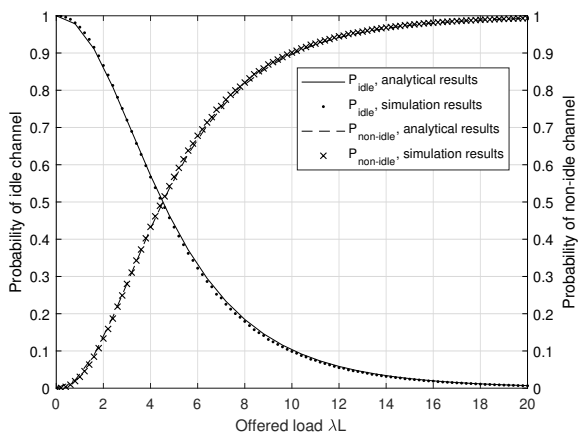


Fig. 7: Probability of idle channel and probability of non-idle channel versus offered load.

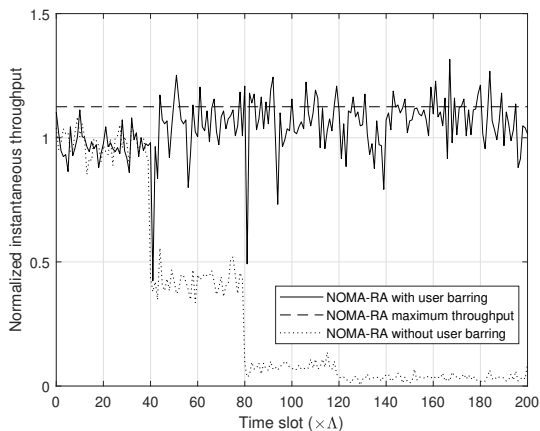


Fig. 8: Actual system throughput versus time slot in an ideal environment.

of 1, gradually decreases, and becomes 0 when the offered load is very heavy. For the probability of non-idle channel, the trend is the opposite. This verifies our initial intention of utilizing the probability of channel being idle/non-idle since it will be a monotonic function with the offered load. Note that the load threshold  $P_{\text{idle}}^{\tau}$  utilized in the user barring algorithm is obtained at the optimal load, i.e.,  $P_{\text{idle}}^{\tau} = P_{\text{idle}}|_{\lambda=\lambda^*}$ . According to the analytical results provided in III-B, it can be calculated

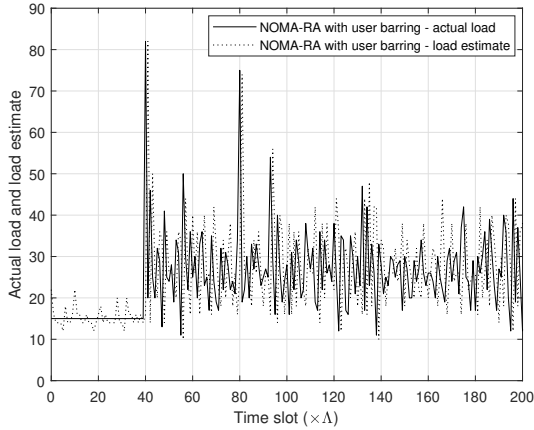
that when  $N = 10$  and  $L = 4$ , the optimal load for the NOMA-RA scheme is 2.6. Then, by using the analytical expressions given in Theorem 4, it can be found that the load threshold  $P_{\text{idle}}^{\tau} = 0.7822$ , which can also be confirmed from Fig. 7.

According to 3GPP technical report for machine-type communications [35], the traffic behavior when a large amount of machine-type devices attempt to access the network after a power outage can be modeled by Beta distribution. In the model, all devices will be activated during a time period  $[0, t]$ , while access density follows the Beta distribution with two parameters  $\alpha = 3$  and  $\beta = 4$ . In our experiment, we assume that there are 1000 devices to be activated within  $10^4$  time slots. By dividing  $10^4$  time slots into 10 periods each occupying  $10^3$  time slots (or  $40\Lambda$  where  $\Lambda = 25$  time slots), the number of devices accessing the system during each period can be calculated by following Beta distribution. To show the response of our proposed user barring algorithm on the increasing loads, in the following simulations, we focus on the first 5 periods, i.e., the first 5000 time slots which captures the increase of traffic loads. For the remaining period, the load gradually decreases as described by the Beta distribution.

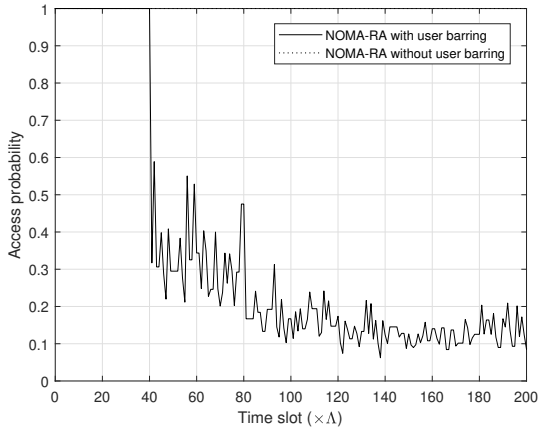
The available resources in the system are given, i.e.,  $N = 10$  and  $L = 4$ . For the considered 5 periods, the number of devices making access attempts can be calculated. The full details are given in Table I. According to Algorithm 1, it is noted that  $P_{\text{access}}$  and  $U_{\text{access}}$  remain fixed during the period of  $\Lambda$ . Within this period, each active user randomly selects one channel and one power level to access in each time slot. After a period of  $\Lambda$ , the normalized instantaneous throughput  $T_{\text{insta}}$  and the relative frequency of idle channels  $P_{\text{idle}}^{\text{insta}}$  are calculated and utilized for load estimate. Hence, to check the performance of the proposed user barring algorithm, it is necessary to check the values of  $T_{\text{insta}}$  over time. Fig. 8 plots the obtained instantaneous throughput for NOMA-RA employed with user barring, in comparison with the throughput achieved without user barring and the maximum throughput. Here, the maximum throughput is obtained by inserting the calculated optimal load  $\lambda^*$  into (6). In Fig. 8, the normalized actual throughput  $T_{\text{insta}}$  versus time slot is shown where the total time period is from 0 to  $200\Lambda - 1$  slots. From Fig. 8, it can be noticed that when the offered load changes from light to

TABLE II: Simulation parameters

Parameter	Value
Cell radius	300 m
Carrier frequency	2 GHz
BS antenna height	15 m
Total occupied bandwidth	1.8 MHz
Bandwidth per channel $B$	180 KHz
Number of channels $N$	10
Total number of users	1000
Number of power levels $L$	4
Target SINR $\Gamma$	6 dB
Noise power spectral density $N_0$	-174 dBm/Hz
Received-power levels	$\gamma_l = \Gamma \sigma^2 (\Gamma + 1)^{L-l}$ , $\forall l = 1, \dots, L$
Transmit power model	Channel inversion [23]
Path loss	3GPP macro propagation model [43]
Shadowing	Log-normal with standard deviation of 10 dB [43]
Fading	Rayleigh flat fading



(a) Actual load estimate



(b) Access probability

Fig. 9: Actual load estimate and access probability versus time slot in an ideal environment.

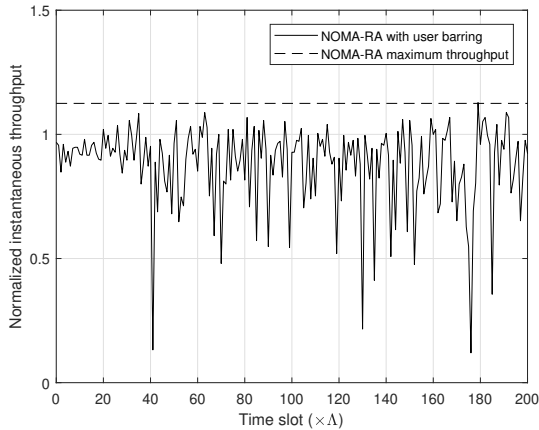
heavy, the normalized throughput achieved by NOMA-RA with user barring stays close to the maximum throughput, while the throughput achieved without user barring decreases dramatically. More specifically, from  $120\Lambda$  onwards, the instantaneous throughput without user barring reduces to almost zero, as the system becomes heavily overloaded. This confirms our design intention and proves the effectiveness of the proposed user barring algorithm for NOMA-RA scheme.

To clearly show the process of load estimate and access probability adjustment, we include Fig. 9(a) and Fig. 9(b) which are plotted for the same system and time period with Fig. 8. Fig. 9(a) plots the actual load in the system, as well as the estimated load. It can be noticed that the load estimate performs very well with excep-

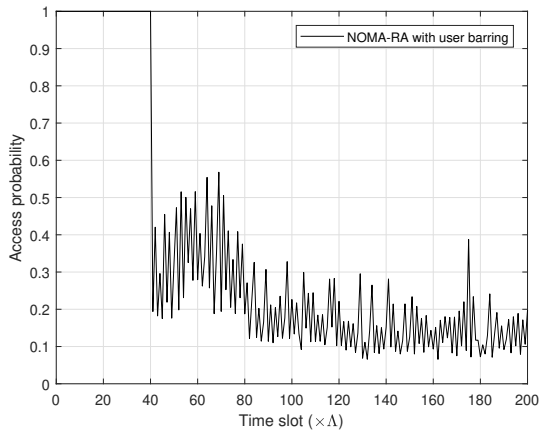
tional levels of accuracy. For the period  $[0, 40\Lambda - 1]$ , the number of active users is always 15, equal to the number of users making access attempts in the system. This is because initially,  $P_{\text{access}}$  is set to be 1 at the time slot 0, which means all 15 users will access. The estimated load, i.e., dotted line in Fig. 9(a), concludes that the system is in light load and therefore  $P_{\text{access}}$  remains to be 1 for the period  $[0, 40\Lambda - 1]$ . For the period  $[40\Lambda, 80\Lambda - 1]$ , the total number of users in the system becomes 83, which is slightly overloaded. If there is no user barring, the access probability  $P_{\text{access}}$  will remain to be 1, as we can notice from Fig. 9(b). Then, the instantaneous throughput will dramatically decrease since the system is continuously overloaded, which can be confirmed from the dotted line in Fig. 8. In contrast, with the proposed user barring algorithm applied, the load can be accurately estimated and the access probability  $P_{\text{access}}$  is adaptively adjusted, as can be observed from Fig. 9(a) and Fig. 9(b). From the time slot  $40\Lambda$  onwards, the total number of users in the system increases in stages. It indeed requires an adaptive load regulation based on the real-time traffic observations which can avoid massive synchronized access demands. Our proposed user barring algorithm eventually results in good performance of instantaneous throughput, as can be seen in Fig. 8.

#### E. User Barring Performance - 3GPP Macro Propagation Model

In this section, we carry out numerical studies for investigating the performance of the proposed user barring algorithm in a more practical wireless propagation



(a) Actual system throughput



(b) Access probability

Fig. 10: Actual system throughput and access probability versus time slot based on a 3GPP macro propagation model.

scenario where the impact of channel variations on signal attenuation is considered, as well as the decoding failure due to the received SINR smaller than a threshold. We consider uniformly distributed users in a cell with a radius of 300 m. Table II summarizes the key parameters used in the simulations. Considering a carrier frequency of 2 GHz and a base station antenna height of 15 m, the path loss (in dB) is calculated using  $128.1 + 37.6 \log_{10}(d)$  [43], where  $d$  is the distance between the base station and the user in kilometers. Then, log-normally distributed shadowing with standard deviation of 10 dB is added to characterize the large-scale propagation effects, together with path loss [43]. Rayleigh flat fading is simulated to reflect the channel variations due to the constructive

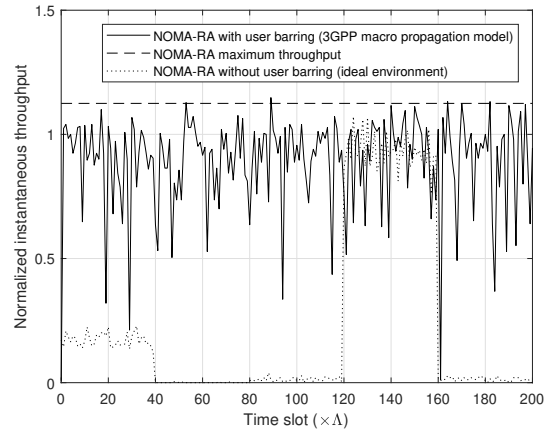


Fig. 11: Actual system throughput versus time slot based on a random number of users in the system, where 3GPP macro propagation model is assumed for barring algorithm and an ideal environment is assumed for no-barring case.

and destructive addition of multipath signal components [44]. Following the same traffic model in Section V-D [35], 1000 devices will be activated during a time period  $[0, 10^4]$ , while access density follows the Beta distribution with two parameters  $\alpha = 3$  and  $\beta = 4$ .

Similar to Section V-D, there are 10 channels and 4 received-power levels in total. The first 5000 time slots, i.e., the first  $200\Lambda$  slots, are taken into account. Fig. 10(a) plots the obtained instantaneous throughput for NOMA-RA with user barring by using the 3GPP propagation model. Although the offered load increases from light to heavy throughout the period  $[0, 200\Lambda - 1]$ , it shows that the instantaneous throughput achieved by NOMA-RA with user barring does not dramatically decrease. By comparing Fig. 10(a) with Fig. 8, we can notice that the achieved system throughput in an ideal environment is larger than that obtained based on a 3GPP propagation model. It is reasonable because in the ideal environment, we assume deterministic channel condition and ignore the decoding failure due to channel variations. Fig. 10(b) plots the access probability, which is adaptively adjusted based on the estimated system load.

Let us recall that the maximum average throughput is achieved at the optimal load  $\lambda^*$ . The proposed user barring algorithm, as shown in previous figures, always guarantees that the obtained instantaneous throughput stays close to the maximum throughput, despite the load variations. To further confirm its effectiveness, Fig. 11 is plotted based on a random number of arriving users in the system. It is noticed that for some time periods,



if there is no user barring, the throughput can be zero. In contrast, the proposed user barring algorithm still provides very good performance, when the offered load level is also random.

## VI. CONCLUSIONS

This paper focused on the novel NOMA-RA scheme and aimed to study the feasibility and benefits of applying it to mMTC scenario in future communication networks. The advantage of NOMA-RA mainly lies in the introduction of power dimension, which allows multiple devices to transmit with the same time-frequency resource, resulting in higher system throughput. Theoretical analysis and comprehensive comparison studies were conducted, which noticed a substantial improvement in throughput. Furthermore, the NOMA-RA scheme is more spectrum-efficient because it showed that only half of the radio resources is required for a system with two pre-determined power levels to achieve a similar performance, compared to the conventional MS-ALOHA with eight channels. Simulation results showed that NOMA-RA is more superior, even when compared to MS-ALOHA with capture effect. This is because, the conventional MS-ALOHA, with or without capture effect, can only support at most one packet per channel in one time slot, while NOMA-RA can support multiple packets. Finally, with the aim of alleviating traffic burstiness in mMTC, we proposed a user barring algorithm which conducts load estimate and continuously adjusts access probability to perform adaptive access control. Simulation results showed that without user barring, the throughput performance dramatically decreases when the offered load changes from light to heavy, while with the proposed user barring algorithm applied, the instantaneous throughput always remains close to the maximum throughput.

As a theoretical investigation, the work presented in this paper assumes perfect physical layer operation such as ideal channel, perfect SIC, and others. In the future work, we shall extend the investigation to consider practical physical layer operation, including the impact of SIC imperfection on the throughput performance of NOMA-RA and our proposed design.

## APPENDIX A

### PROOF FOR THEOREM 1

Assume that there are  $U_i$  packets to be transmitted on a channel, say, channel  $i$ . Within a given time slot, each packet randomly chooses one power level to access, where power collisions may happen. In order to calculate the average number of successful packets over  $L$  power

levels on channel  $i$ , we first calculate the conditional probability of having  $S_i > 0$  packets successful, i.e.,  $P(S_i|U_i)$ , where the number of successful packets  $S_i$  falls in the range of  $[1, \min(U_i, L)]$ . We exclude the case  $S_i = 0$  since this case produces no successful packet and can be omitted when computing the throughput.

To calculate  $P(S_i|U_i)$ , we consider two cases where the first case includes the presence of power collision and the second case does not contain any power collision. In the first case, we have the condition  $S_i < \min(L, U_i)$  indicating that the number of successful packets is less than the number of contending packets and the number of power levels. To derive  $P(S_i|U_i)$ , we first identify the level where power collision first occurs, then we derive the probability that there are exactly  $S_i$  successful packets appearing above the power collision level. If the power collision appears at the  $(S_i + g + 1)^{\text{th}}$  power level, this indicates that there are  $g$  idle power levels appearing above the power collision level. The probability,  $P(S_{i,g}|U_i)$ , that there are exactly  $S_i$  successful packets and  $g$  idle power levels appearing above the power collision level can be computed by

$$\frac{\binom{U_i}{S_i} S_i! \binom{S_i+g}{S_i}}{L^{U_i}} \left( (L - S_i - g)^{U_i - S_i} - (L - S_i - g - 1)^{U_i - S_i} - \binom{U_i - S_i}{1} (L - S_i - g - 1)^{U_i - S_i - 1} \right), \quad (11)$$

where  $\binom{n}{r}$  is the binomial coefficient. Then, the conditional probability  $P(S_i|U_i)$  can be given by taking into account all the possible cases of  $g$ , i.e.,  $P(S_i|U_i) = \sum_{g=0}^{L-1-S_i} P(S_{i,g}|U_i)$ .

In the second case, we have  $S_i = U_i \leq L$  where all  $U_i$  packets are successful. The successful probability  $P(S_i|U_i)$  can be calculated simply by  $\frac{\binom{U_i}{S_i} S_i! \binom{L}{S_i}}{L^{U_i}}$ . Further, for any other cases satisfying  $S_i \in [1, \min(U_i, L)]$  but do not belong to the above scenarios, we have  $P(S_i|U_i) = 0$ . Finally, we obtain the expression for  $P(S_i|U_i)$ , given in (4).

We now derive the conditional throughput which is defined as the average number of successfully transmitted packets on channel  $i$ . Given that there are  $U_i$  packets sent to access channel  $i$ , the conditional throughput  $E[S_i|U_i]$  can be determined by  $\sum_{S_i=1}^{\min(L, U_i)} S_i P(S_i|U_i)$ , by taking into account all possible numbers of successful packets.

## APPENDIX B

### PROOF FOR THEOREM 3

Let us focus on a single channel with  $L$  power levels. Each power level on one channel can be considered as one virtual resource block, so there are  $L$  ordered virtual resource blocks for one channel. Assume that the packet arrival on each power level follows Poisson distribution with a parameter  $\lambda$ . In other words, the probability of  $k$  packets arriving at each power level is given by  $q_k = \frac{\lambda^k e^{-\lambda}}{k!}$ , where  $k \in \{0, 1, 2, \dots\}$ . The probability of having a successful packet on the  $i^{\text{th}}$  power level (counting from the highest) is simply the probability that a successful packet appears at the  $i^{\text{th}}$  power level (which is  $q_1$ ) and no power collision appears at all the above power level (which is  $(q_0 + q_1)^{i-1}$ ). Since each power level can carry a successful packet, the number of successful packets a channel can produce is thus  $\sum_{i=1}^L q_1 (q_0 + q_1)^{i-1}$  or  $\sum_{i=1}^L \lambda e^{-\lambda} (e^{-\lambda} + \lambda e^{-\lambda})^{i-1}$ . When multiple channels are considered, the total throughput  $T$  for  $N$  channels can be computed by  $N \sum_{i=1}^L \lambda e^{-\lambda} (e^{-\lambda} + \lambda e^{-\lambda})^{i-1}$ , due to the fact that all channels are independently and uniformly accessed by users.

## APPENDIX C

### PROOF FOR THEOREM 4

Let us focus on a particular channel. Given a Poisson arrival with a rate of  $\lambda$ , the probability of  $k$  packets arriving at each power level is given by  $q_k = \frac{\lambda^k e^{-\lambda}}{k!}$ , where  $k \in \{0, 1, 2, \dots\}$ . Recall that an "idle channel" is labeled if any idle power level is observed before a power collision or before all power levels are successfully decoded for the case of no power collision<sup>12</sup>. In the following, we shall derive the idle channel probability.

The probability of having a power collision appearing at a particular level, say, the  $i^{\text{th}}$  power level, is  $1 - q_0 - q_1$ . To ensure this power collision is the first appearance on the channel counting from the highest power level, all the higher  $i - 1$  power levels must not contain a power collision. The probability that a power level does not contain a power collision is  $q_0 + q_1$ , and for all the power levels above the  $i^{\text{th}}$  to happen, the probability is  $(q_0 + q_1)^{i-1}$ . However, this event includes the case that no idle power level exists above the power collision which should be

<sup>12</sup>Please be reminded that a positive number of arrivals are assumed, because when there is no arriving packet, the probability of idle channel is simply one.

excluded. This case can happen when each power level above the  $i^{\text{th}}$  power level contains a successful packet, with a probability of  $(q_1)^{i-1}$ . By excluding this case, we have  $(q_0 + q_1)^{i-1} - (q_1)^{i-1}$ . Hence, the probability of the first power collision happening at the  $i^{\text{th}}$  power level is given by

$$(1 - q_0 - q_1)((q_0 + q_1)^{i-1} - (q_1)^{i-1}). \quad (12)$$

Finally, considering that the first power collision can occur in any power level except the first which does not allow any idle power level, we get the probability of idle channel, given as

$$(1 - q_0 - q_1) \left( \sum_{i=2}^L (q_0 + q_1)^{i-1} - (q_1)^{i-1} \right), \quad (13)$$

which gives

$$(1 - e^{-\lambda} - \lambda e^{-\lambda}) \left( \sum_{i=2}^L (e^{-\lambda} + \lambda e^{-\lambda})^{i-1} - (\lambda e^{-\lambda})^{i-1} \right). \quad (14)$$

The above forms the last term in (10).

On the other hand, it is possible that there is at least an idle power level but no power collision appearing on the channel. In other words, all power levels must contain either zero or one packet transmission (with a probability of  $(q_0 + q_1)^L$ ), and we need to exclude the case where each power level is occupied by exactly one packet transmission which creates no idle power level on the channel (with probability of  $(q_1)^L$ ). With the above, we get the probability of idle channel for this scenario, given by

$$(e^{-\lambda} + \lambda e^{-\lambda})^L - (\lambda e^{-\lambda})^L, \quad (15)$$

which forms the first two terms in (10). This completes the proof of (10) in Theorem 4.

## REFERENCES

- [1] Cisco Visual Networking Index, "Global mobile data traffic forecast update, 2017-2022," Cisco, San Jose, CA, USA, Tech. Rep., Feb. 2019.
- [2] C. Bockelmann, N. K. Pratas, G. Wunder *et al.*, "Towards massive connectivity support for scalable mMTC communications in 5G networks," *IEEE Access*, vol. 6, pp. 28 969–28 992, May 2018.
- [3] S. Moon, H. Lee, and J. Lee, "SARA: Sparse code multiple access-applied random access for IoT devices," *IEEE Internet Things J.*, vol. 5, no. 4, pp. 3160–3174, Aug. 2018.
- [4] 3GPP TS 36.211, "Evolved universal terrestrial radio access (E-UTRA) physical channels and modulation," 3GPP, Tech. Rep., Release 13, 2016.
- [5] L. Feltrin, G. Tsoukaneri, M. Condoluci *et al.*, "Narrowband IoT: A survey on downlink and uplink perspectives," *IEEE Wireless Commun.*, vol. 26, no. 1, pp. 78–86, Feb. 2019.
- [6] N. Jiang, Y. Deng, M. Condoluci, W. Guo, A. Nallanathan, and M. Dohler, "RACH preamble repetition in NB-IoT network," *IEEE Commun. Lett.*, vol. 22, no. 6, pp. 1244–1247, Jun. 2018.

- [7] N. Jiang, Y. Deng, A. Nallanathan, and J. A. Chambers, "Reinforcement learning for real-time optimization in NB-IoT networks," *IEEE J. Sel. Areas Commun.*, vol. 37, no. 6, pp. 1424–1440, Jun. 2019.
- [8] Y. Liu, Z. Qin, M. ElKashlan, Z. Ding, A. Nallanathan, and L. Hanzo, "Non-orthogonal multiple access for 5G and beyond," *Proceedings of the IEEE*, vol. 105, no. 12, pp. 2347–2381, Dec. 2017.
- [9] Z. Ding, Y. Liu, J. Choi, Q. Sun, M. ElKashlan, I. Chih-Lin, and H. V. Poor, "Application of non-orthogonal multiple access in LTE and 5G networks," *IEEE Commun. Mag.*, vol. 55, no. 2, pp. 185–191, Feb. 2017.
- [10] M. B. Shahab, R. Abbas, and S. J. J. M. Shirvanimoghadam, "Grant-free non-orthogonal multiple access for IoT: A survey," *IEEE Commun. Surveys Tuts.*, May in press, 2020.
- [11] W. Yu, A. Chorti, L. Musavian, H. V. Poor, and Q. Ni, "Effective secrecy rate for a downlink NOMA network," *IEEE Trans. Wireless Commun.*, in press, 2019.
- [12] L. Miuccio, D. Panno, and S. Riolo, "Joint control of random access and dynamic uplink resource dimensioning for massive mtc in 5G NR based on SCMA," *IEEE Internet Things J.*, vol. 7, no. 6, pp. 5042–5063, Jun. 2020.
- [13] M. Shirvanimoghadam, M. Condoluci, M. Dohler, and S. J. Johnson, "On the fundamental limits of random non-orthogonal multiple access in cellular massive IoT," *IEEE J. Sel. Areas Commun.*, vol. 35, no. 10, pp. 2238–2252, Oct. 2017.
- [14] J. Ding, D. Qu, and J. Choi, "Analysis of non-orthogonal sequences for grant-free ra with massive MIMO," *IEEE Trans. Commun.*, vol. 68, no. 1, pp. 150–160, Jan. 2020.
- [15] E. Balevi, F. T. A. Rabee, and R. D. Gitlin, "ALOHA-NOMA for massive machine-to-machine IoT communication," in *Proc. IEEE Int. Conf. Commun. (ICC)*, Kansas City, MO, USA, May 2018.
- [16] S. A. Tegos, P. D. Diamantoulakis, A. S. Lioumpas, P. G. Sarigiannidis, and G. K. Karagiannidis, "Slotted ALOHA with NOMA for the next generation IoT," *IEEE Trans. Commun.*, Jul. in press, 2020.
- [17] Z. Chen, Y. Liu, S. Khairy, L. X. Cai, Y. Cheng, and R. Zhang, "Optimizing non-orthogonal multiple access in random access networks," in *Proc. IEEE Veh. Technol. Conf. (VTC)*, Antwerp, Belgium, Belgium, May 2020.
- [18] Y. Wang, T. Wang, Z. Yang, D. Wang, and J. Cheng, "Throughput-oriented non-orthogonal random access scheme for massive MTC networks," *IEEE Trans. Commun.*, vol. 68, no. 3, pp. 1777–1793, Mar. 2020.
- [19] Y. Liang, X. Li, J. Zhang, and Z. Ding, "Non-orthogonal random access for 5G networks," *IEEE Trans. Wireless Commun.*, vol. 16, no. 7, pp. 4817–4831, Jul. 2017.
- [20] J.-B. Seo, H. Jin, and B. C. Jung, "Non-orthogonal random access with channel inversion for 5G networks," in *Proc. Int. Conf. Inf. Commun. Technol. Convergence*, Jeju, South Korea, Oct. 2017.
- [21] J.-B. Seo, B. C. Jung, and H. Jin, "Performance analysis of NOMA random access," *IEEE Commun. Lett.*, vol. 22, no. 11, pp. 2242–2245, Nov. 2018.
- [22] S. M. R. Islam, N. Avazov, O. A. Dobre, and K. Kwak, "Power-domain non-orthogonal multiple access (NOMA) in 5G systems: Potentials and challenges," *IEEE Commun. Surveys Tuts.*, vol. 19, no. 2, pp. 721–742, Second Quarter 2017.
- [23] J. Choi, "NOMA-based random access with multi-channel ALOHA," *IEEE J. Sel. Areas Commun.*, vol. 35, no. 12, pp. 2736–2743, Dec. 2017.
- [24] H. S. Jang, H. Lee, and T. Q. S. Quek, "Deep learning-based power control for non-orthogonal random access," *IEEE Communications Letters*, vol. 23, no. 11, pp. 2004–2007, 2019.
- [25] M. V. da Silva, R. D. Souza, H. Alves, and T. Abrão, "A noma-based q-learning random access method for machine type communications," *IEEE Wireless Communications Letters*, in press, 2020.
- [26] J. Choi, "Layered non-orthogonal random access with SIC and transmit diversity for reliable transmissions," *IEEE Trans. Commun.*, vol. 66, no. 3, pp. 1262–1272, Mar. 2018.
- [27] R. O. LaMaire and A. Krishna, "Maximizing throughput in a random access radio system by optimal power level choice," in *Proc. IEEE Int. Conf. Commun. (ICC)*, Dallas, TX, USA, Jun. 1996, pp. 614–620.
- [28] A. S. Tanenbaum and D. J. Wetherall, *Computer Networks*, 5th ed. Pearson, Sep. 2010.
- [29] D. G. Jeong and W. S. Jeon, "Performance of an exponential backoff scheme for slotted-ALOHA protocol in local wireless environment," *IEEE Trans. Veh. Technol.*, vol. 44, no. 3, pp. 470–479, Aug. 1995.
- [30] 3GPP TS 22.011, "Technical specification group services and system aspects," 3GPP, Tech. Rep., Release 13, 2016.
- [31] M. Grau, C. H. Foh, A. ul Quddus, and R. Tafazolli, "Preamble barring: A novel random access scheme for machine type communications with unpredictable traffic bursts," in *Proc. IEEE Veh. Technol. Conf. (VTC)*, Hawaii, USA, Sep. 2019.
- [32] L. P. Qian, Y. Wu, H. Zhou, and X. Shen, "Joint uplink base station association and power control for small-cell networks with non-orthogonal multiple access," *IEEE Trans. Wireless Commun.*, vol. 16, no. 9, pp. 5567–5582, 2017.
- [33] K. Wang, Y. Liu, Z. Ding, A. Nallanathan, and M. Peng, "User association and power allocation for multi-cell non-orthogonal multiple access networks," *IEEE Trans. Wireless Commun.*, vol. 18, no. 11, pp. 5284–5298, 2019.
- [34] L. Tello-Quendo, I. Leyva-Mayorga, V. Pla, J. Martinez-Bauset, J. Vidal, V. Casares-Giner, and L. Guijarro, "Performance analysis and optimal access class barring parameter configuration in lte-a networks with massive m2m traffic," *IEEE Trans. Veh. Technol.*, vol. 67, no. 4, pp. 3505–3520, 2018.
- [35] 3GPP TR 37.868, "Study on RAN improvements for machine-type communications." Technical report, Tech. Rep., Release 11, 2011.
- [36] T. L. Marzetta and B. M. Hochwald, "Fast transfer of channel state information in wireless systems," *IEEE Transactions on Signal Processing*, vol. 54, no. 4, pp. 1268–1278, 2006.
- [37] G. Caire, N. Jindal, M. Kobayashi, and N. Ravindran, "Multiuser MIMO achievable rates with downlink training and channel state feedback," *IEEE Transactions on Information Theory*, vol. 56, no. 6, pp. 2845–2866, 2010.
- [38] 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, Oct. 2018.
- [39] X. Li, H. Liu, S. Roy, J. Zhang, P. Zhang, and C. Ghosh, "Throughput analysis for a multi-user, multi-channel ALOHA cognitive radio system," *IEEE Trans. Wireless Commun.*, vol. 11, no. 11, pp. 3900–3909, Nov. 2012.
- [40] H. Zhou and R. H. Deng, "Capture model for mobile radio slotted ALOHA systems," *IEE Proc. Commun.*, vol. 145, no. 2, pp. 91–97, Apr. 1998.
- [41] J. F. Kurose and K. W. Ross, *Computer Networking : A Top-Down Approach*, 6th ed. Pearson, May 2012.
- [42] M. Vilgelm, S. Schiessl, H. Al-Zubaidy, W. Kellerer, and J. Gross, "On the reliability of LTE random access: Performance bounds for machine-to-machine burst resolution time," in *Proc. IEEE Int. Conf. Commun. (ICC)*, Kansas City, MO, USA, May 2018.
- [43] 3GPP TR 25.942, "Universal mobile telecommunications system (UMTS); radio frequency (RF) system scenarios," Technical report, Tech. Rep., Release 14, 2017.
- [44] A. J. Goldsmith, *Wireless Communications*. Cambridge: Cambridge Univ. Press, 2005.