Multiple Access for Massive Machine Type Communications

Rana Abbas

A thesis submitted in fulfillment
of the requirements of the degree of
Doctor of Philosophy

THE UNIVERSITY OF
SYDNEY

Australian Centre for Excellence in Telecommunications
School of Electrical and Information Engineering
The University of Sydney

October 2017

Declaration

I hereby declare that this submission is my own work and that, to the best of my knowledge and belief, it contains no material previously published or written by another person nor material which to a substantial extent has been accepted for the award of any other degree or diploma of the University or other institute of higher learning, except where due acknowledgement has been made in the text.

Rana Abbas

11 April 2018

Abstract

The internet we have known thus far has been an internet of people, as it has connected people with one another. However, these connections are forecasted to occupy only a minuscule of future communications. The internet of tomorrow is indeed: the internet of things. The Internet of Things (IoT) promises to improve all aspects of life by connecting everything to everything. An enormous amount of effort is being exerted to turn these visions into a reality. Sensors and actuators will communicate and operate in an automated fashion with no or minimal human intervention. In the current literature, these sensors and actuators are referred to as machines, and the communication amongst these machines is referred to as Machine to Machine (M2M) communication or Machine-Type Communication (MTC). As IoT requires a seamless mode of communication that is available anywhere and anytime, wireless communications will be one of the key enabling technologies for IoT.

In existing wireless cellular networks, users with data to transmit first need to request channel access. All access requests are processed by a central unit that in return either grants or denies the access request. Once granted access, users' data transmissions are non-overlapping and interference free. However, as the number of IoT devices is forecasted to be in the order of hundreds of millions, if not billions, in the near future, the access channels of existing cellular networks are predicted to suffer from severe congestion and, thus, incur unpredictable latencies in the system. On the other hand, in random access, users with data to transmit will access the channel in an uncoordinated and probabilistic fashion, thus, requiring little or no signalling overhead. However, this reduction in overhead is at the expense of reliability and efficiency due to the interference caused by contending users. In most existing random access schemes, packets are lost when they experience interference from other packets transmitted over the same resources. Moreover, most existing random access schemes are best-effort schemes with almost no Quality of Service (QoS) guarantees. In this thesis, we investigate the performance of different random access schemes in different settings to resolve the problem of the massive access of IoT devices with diverse QoS guarantees.

First, we take a step towards re-designing existing random access protocols such that they are more practical and more efficient. For many years, researchers have

adopted the collision channel model in random access schemes: a collision is the event of two or more users transmitting over the same time-frequency resources. In the event of a collision, all the involved data is lost, and users need to retransmit their information. However, in practice, data can be recovered even in the presence of interference provided that the power of the signal is sufficiently larger than the power of the noise and the power of the interference. Based on this, we re-define the event of collision as the event of the interference power exceeding a pre-determined threshold. We propose a new analytical framework to compute the probability of packet recovery failure inspired by error control codes on graph. We optimize the random access parameters based on evolution strategies. Our results show a significant improvement in performance in terms of reliability and efficiency.

Next, we focus on supporting the heterogeneous IoT applications and accommodating their diverse latency and reliability requirements in a unified access scheme. We propose a multi-stage approach where each group of applications transmits in different stages with different probabilities. We propose a new analytical framework to compute the probability of packet recovery failure for each group in each stage. We also optimize the random access parameters using evolution strategies. Our results show that our proposed scheme can outperform coordinated access schemes of existing cellular networks when the number of users is very large.

Finally, we investigate random non-orthogonal multiple access schemes that are known to achieve a higher spectrum efficiency and are known to support higher loads. In our proposed scheme, user detection and channel estimation are carried out via pilot sequences that are transmitted simultaneously with the user's data. Here, a collision event is defined as the event of two or more users selecting the same pilot sequence. All collisions are regarded as interference to the remaining users. We first study the distribution of the interference power and derive its expression. Then, we use this expression to derive simple yet accurate analytical bounds on the throughput and outage probability of the proposed scheme. We consider both joint decoding as well as successive interference cancellation. We show that the proposed scheme is especially useful in the case of short packet transmissions.

To my beloved parents,

Abdulwahed and Doha

Acknowledgements

I would like to acknowledge the help and support of my two supervisors: Prof. Yonghui Li and Prof. Branka Vucetic. Prof. Yonghui has been unwaveringly supportive, helpful, knowledgeable, resourceful, and very kind during my PhD studies. Prof. Branka Vucetic has been a role model and a true inspiration. I am grateful to have had the privilege of being her student and research assistant over the past few years. I would like to thank her immensely for the opportunity and for always taking the time to give me constructive advice on my research as well as career.

Dr. Mahyar Shirvanimoghaddam has played a fundamental role in my research and my PhD as a whole. He has been most patient, most helpful, a genuine supporter, colleague and friend. My sincerest gratitude goes out to him for all that he has done for me. Amongst the many others, who have helped me throughout this journey and have generously given me a significant amount of their time and advice, I would like to thank Dr. He Chen, Dr. Zihuai Lin, Dr. Jun Li, Dr. Peng Wang, Dr. Wibowo Hardjwanna, and Dr. Wei Bao. I would also like to thank my friends at the Centre of Excellence in Telecommunications for making the working environment and the journey as a whole a pleasant one.

I would also like to thank my various sources of financial support: the Australian Research Council (ARC) for the APA Scholarship, the University of Sydney for the PRSS scheme, the school of Electrical and Information Engineering at the University of Sydney for the Norman I Prize scholarship, and Prof. Yonghui Li for his generous funding of my conference trips.

I am indebted to all my family and friends as they were my supporters through my PhD, and they will forever be my supporters through life. My uncle Wahib and his wife Zena, I thank you for taking me into your loving home during my studies in Australia. Nothing I say or do will ever be enough repayment. To my siblings Mohamad, Khaled, Yahya and Nour, I simply cannot do without you. Finally, to my husband Rasheed that has stood by me throughout the ups and downs of PhD, and to my parents Abdulwahed and Doha: *You are my all!*.

At last, I thank God for all that he has blessed me with. I will forever continue to pray for knowledge that is of benefit, a good provision and deeds that will be accepted.

Contents

List of Figures

List of Tables

List of Acronyms

3GPP	Third Generation Partnership Project
ACB	Access Class Barring
AP	Access Point
AWGN	Additive White Gaussian Noise
BEC	Binary Erasure Channel
BS	Base Station
CDF	Cumulative Density Function
CDMA	Code Division Multiple Access
CF	Characteristic Function
CMA-ES	Covariance Matrix Adaptation Evolution Strategy
CRN	Cognitive Radio Network
CSI	Channel State Information
CSMA	Carrier Sense Multiple Access
CSMA-CA	Carrier Sense Multiple Access with Collision Avoidance
D2D	Device to Device
DAB	Dynamic Access Barring
DAMA	Demand Assigned Multiple Access
EAB	Extended Access Barring
FDMA	Frequency Division Multiple Access
H2H	Human to Human
IoT	Internet of Things
IIoT	Industrial Internet of Things
IP	Interference Power
IRSA	Irregular Repeat Slotted ALOHA
LDPC	Low Density Parity Check
LDS	Low-Density Spreading
LT	Luby Transform
LTE	Long Term Evolution - 4G
LTE-A	Long Term Evolution Advanced
M2M	Machine to Machine
MAC	Multiple Access Channel
mMTC	massive Machine-Type Communication
ML	Maximum Likelihood

MRC	Maximal Ratio Combining
MTC	Machine-Type Communication
MUSA	Multiple User Spectrum Access
MUST	Multi User Superposition Transmission
NOMA	Non-Orthogonal Multiple Access
OFDM	Orthogonal Frequency Division Multiplexing
OFDMA	Orthogonal Frequency Division Multiple Access
OMA	Orthogonal Multiple Access
PDF	Probability Density Function
PDMA	Pattern Division Multiple Access
PPP	Poisson Point Process
PRACH	Physical Random Access Channel
PU	Primary User
QoS	Quality of Service
RA	Random Access
RACH	Random Access Channel
RB	Resource Block
S-ALOHA	Slotted ALOHA
SCMA	Sparse Code Multiple Access
SCFDMA	Single Carrier Frequency Division Multiple Access
SDMA	Spatial Division Multiple Access
SIC	Successive Interference Cancellation
SINR	Signal-to-Interference and Noise Ratio
SJD	Successive Joint Decoding
SNR	Signal-to-Noise Ratio
SoDeMA	Software Defined Multiple Access
SU	Secondary User
TDMA	Time Division Multiple Access
UE	User Equipment
URLLC	Ultra-high Reliability Low Latency Communication
WCDMA	Wide-band Code Division Multiple Access

List of Related Works

The following is a list of submitted and published papers in refereed journals and conference proceedings produced during my Ph.D candidature. In some cases, the conference papers contain material overlapping with the journal papers.

Journal Papers

[J1] Abbas, Rana, Chen, He, Yonghui Li, and Branka Vucetic. "Scheduling for Short Packets." *To be submitted to IEEE Communication Letters.*

[J2] Abbas, Rana, Mahyar Shirvanimoghaddam, Yonghui Li, and Branka Vucetic. "Grant-Free Massive NOMA: Outage Probability and Throughput." arXiv preprint arXiv:1707.07401 (2017). *Submitted to IEEE Transactions on Communications.*

[J3] Abbas, Rana, Mahyar Shirvanimoghaddam, Yonghui Li, and Branka Vucetic. "Random Access for M2M Communications With QoS Guarantees." IEEE Transactions on Communications 65, no. 7 (2017): 2889-2903.

Conference Papers

[C1] Abbas, Rana, Mahyar Shirvanimoghaddam, Yonghui Li, and Branka Vucetic. "On the Performance of Grant-Free Massive NOMA." Accepted to appear in the Proceedings of IEEE 28th Annual International Symposium on Personal, Indoor, and Mobile Radio Communications (PIMRC), 2017. *Winner of Best Paper Award.*

[C2] Abbas, Rana, Mahyar Shirvanimoghaddam, Yonghui Li, and Branka Vucetic. "Analysis on LT codes for unequal recovery time with complete and partial feedback." In the Proceedings of IEEE International Symposium in Information Theory (ISIT), pp. 305-309, 2016.

[C3] Abbas, Rana, Mahyar Shirvanimoghaddam, Yonghui Li, and Branka Vucetic. "Performance analysis and optimization of LT codes with unequal recovery time and intermediate feedback." In the Proceedings of IEEE International Conference in Communications (ICC), pp. 1-6, 2016.

[C4] Abbas, Rana, Mahyar Shirvanimoghaddam, Yonghui Li, and Branka Vucetic. "On SINR-Based Random Multiple Access Using Codes on Graph." In the Proceedings of IEEE Global Communications Conference (GLOBECOM), pp. 1-6, 2015.

[C5] Abbas, Rana, Mahyar Shirvanimoghaddam, Yonghui Li, and Branka Vucetic. "Design of probabilistic random access in cognitive radio networks." In the Proceedings of International Conference on Cognitive Radio Oriented Wireless Networks (CROWNCOM), pp. 696-707. Springer, Cham, 2015.

Magazine Papers

[M1] Chen, He, Rana Abbas, Peng Cheng, Mahyar Shirvanimoghaddam, Wibowo Hardjawana, Wei Bao, Yonghui Li, and Branka Vucetic. "Ultra-Reliable Low Latency Cellular Networks: Use Cases, Challenges and Approaches." arXiv preprint arXiv:1709.00560 (2017). *Submitted to IEEE Network - The Magazine of Global Internetworking- on 5G for Ultra-Reliable Low Latency Communications.*

[M2] Jiao, Jian, Rana Abbas, Yonghui Li, and Qinyu Zhang. "Multiple Access Rateless Network Coding for Machine-to-Machine Communications." ZTE Communications 35 (2016): 1.

[M3] Mahyar Shirvanimoghaddam, Mohamad Sadegh Mohamadi, Rana Abbas, Aleksandar Minja, Balazs Matuz, Guojun Han, Zihuai Lin, Yonghui Li, Sarah Johnson, and Branka Vucetic. "Short Block-length Codes for Ultra-Reliable Low-Latency Communications." arXiv preprint arXiv:1802.09166 (2018).*Submitted to IEEE Communications Magazine.*

Chapter 1

Introduction

1.1 Internet of Things

"*IoT enables physical objects to see, hear, think and perform jobs by having them "talk" together, to share information and to coordinate decisions. The IoT transforms these objects from being traditional to smart by exploiting its underlying technologies such as ubiquitous and pervasive computing, embedded devices, communication technologies, sensor networks, Internet protocols and applications,*"[1].

This is one of the many definitions of IoT that can be found in the research community and in the industry [2, 3, 4, 5, 6]. The existing large number of definitions, albeit confusing at times, emphasizes the importance of IoT on many levels [4, 7, 8, 9, 10, 11, 12, 1, 13]. The term was first coined in the context of the management of supply chains back in 1999 [14]. Since then, IoT has been generalized to cover a vast range of applications. Although its range of applications continues to grow and its enabling technologies continue to evolve, the objective remains to improve the quality of life, society and industry as well as their productivity. IoT will achieve these objectives by connecting all the different related systems to one another so that decision-making policies and actions are based on the integrated data and, thus, are more effective. In what follows, some of the main applications of IoT are reviewed.

- *Smart Home*: Smart Homes allow the customization and control of homes via a touch panel and, consequently, provide means to reduce costs, save time, increase security and improve energy efficiency. Some of the technologies available include: smart electricity and water meters, smart refrigerators, smart garden irrigation systems, automated heating, ventilation and air conditioning systems, automated indoor and outdoor lighting, automated sound and security systems, etc.

- *Smart City*: Smart Cities will tackle major problems such as the growing and aging populations, climate changes, pollution, traffic congestion, energy supply shortage, etc. Use cases cover smart surveillance and maintenance, smart energy management systems, smart water distribution, environmental monitoring, traffic congestion control, waste management, etc.

- *Intelligent Transport Systems*: The first generation of Intelligent Transportation Systems (ITS) included stand-alone systems. However, future ITS will require the cooperation of these systems and the sharing of data for better decision making policies and, consequently, improved efficiency, safety, durability, environmental impact and overall in-car experience. Thus, the future generation of ITS is now being referred to as Cooperative Intelligent Transportation Systems (C-ITS).

- *Industrial Applications*: One application of Industrial Internet of Things (IIoT) is the reliable prediction of onshore and offshore pump failures in oil and gas mining industries to reduce production losses. IIoT will also allow the reliable detection of sources of power outages and faults in Smart Grids. More general use cases involve remote monitoring of large scale areas (that are too costly to survey manually), reduction of factory carbon emissions, fast and reliable prediction, detection and response to critical events, etc.

- *Health care and Medicine*: IoT will enable the tracking and monitoring of all related objects in the health care system, i.e., patients, doctors, hospitals, equipment, medicine, etc. The collection of information such as temperature, pres-

sure, daily activity, and the availability of this information to the doctors and laboratories at any given time or place will allow remote consultation, diagnosis, therapy, medication, and even surgery.

1.2 Research Problems and Contributions

No doubt, there are many challenges facing the realization and evolution of IoT. To name a few, these challenges include device identification, sensing and actuation, computational and processing capabilities, device power, manufacturing cost, communications and networking, security, etc. This thesis focuses on the wireless communications part of the system. Both the research and industry communities agree that wireless communications will be a key component of IoT, especially when involving mobility, deployment in remote areas, high temperature or pressure environments, underground, underwater, etc. In such environments, wired communications is either too costly or not possible at all.

In particular, this thesis focuses on the wireless access of the involved devices, which was highlighted by the Third Generation Partnership Project (3GPP) as one of the major challenges that are hindering the progress of M2M. The main idea is that prior to data transmission, users need to request access to the available time-frequency resources. Coordinated access relies on a central processing unit that controls the time and frequency resource allocation based on either the received requests or on polling.

It is clear that the applications of IoT are heterogeneous both in nature and requirements. Therefore, one access scheme will not suffice to satisfy all requirements. In fact, these applications have been categorized into two main streams. The first stream encompasses mission critical applications that require Ultra-high Reliability Low Latency Communication (URLLC), e.g., robotic surgery, critical safety messages in ITS, etc. On the other hand, the second stream encompasses applications with less stringent requirements on reliability and latency, e.g., smart metering. The main challenge for these applications is scalability: the support of the massive number

of devices while guaranteeing satisfactory performances. Thus, it is referred to as massive Machine-Type Communication (mMTC). In this thesis, we focus on mMTC.

For mMTC, we can summarize the main research problems into four categories. The first category is related to the large number of devices and the sporadic nature of their traffic which is predicted to lead to severe congestion at the access channels; consequently, cellular networks will suffer from unpredictable delays, increased energy expenditure due to the increased number of retransmissions and increased processing and computational complexity at the central unit. The second category is related to the small payloads of these devices that are often smaller than the signalling overhead required in cellular networks; thus, the current access protocols are inefficient. The third category is related to the heterogeneity of the applications and use cases covered, which lead to a large variation in the delay and reliability requirements. Finally, the forth category is in relation to the low power budget of these devices which requires them to have low duty cycles, low transmission powers and low computational and processing capabilities. Based on all of the above, in particular categories one and two, there is large consensus among research and industry partners that uncoordinated/random access is more suitable than coordinated access, and, for that, we focus on the analysis, design and optimization of random wireless access in this thesis to satisfy the delay, reliability and power requirements of mMTC. The main contributions are listed below.

1.2.1 The Analysis and Design of Practical Random Access for Higher System Throughput

Most works on random access consider the collision model where a collision is defined as the event of two or more users transmitting over the same time-frequency resources. A collision leads to the loss of these users' information, and users usually need to retransmit their information. However, this model treats interference as a binary parameter. In the absence of interference, a packet is said to be clean and successfully recovered with high probability. Otherwise, it is corrupted and lost.

In practice, interference is the accumulation of the power received from other users that are concurrently transmitting, which can never be zero. For that, the Signal-to-Interference and Noise Ratio (SINR) is a more accurate metric for analyzing and designing random access schemes [15], as it distinguishes between different levels of interference. Thus, a more practical definition of collision is the event of having a user's SINR drop below a predefined threshold. An SINR-based random access model can support higher system loads subject to the same power and performance requirements. Based on the above, the *first* research problem we tackle in this thesis (Chapter 3) is the analysis and design of SINR-based Random Access (RA) using Successive Interference Cancellation (SIC) (ref. [C5] and [C4]).

1. The packet recovery process in conventional RA using SIC was shown to be analogous to the iterative recovery process of codes-on-graph for the Binary Erasure Channel (BEC), thus, allowing the direct application of the AND-OR tree analysis. In this chapter, we first extend the AND-OR tree analysis to a more generalized framework that can track the evolution of error probabilities in each iteration of the SIC process based on the SINR metric.

2. Based on the derived analytical framework, we optimize the design parameters using evolution strategies such that the error probabilities are minimized. Our numerical results show that the SINR metric allows for the support of a larger number of devices, under the same power constraints and reliability requirements.

3. We apply the proposed SINR-based RA to Cognitive Radio Networks (CRNs), where the Secondary Users (SUs) can reuse the resources of the Primary Users (PUs), provided that the QoS of the primary network is protected. We formulate and solve a new optimization problem to find the optimal transmission scheme that maximizes the throughput of the SU network. Our proposed design can achieve higher throughput and is more energy efficient than conventional schemes.

1.2.2 The Analysis and Design of Random Access with Diverse Quality of Service Requirements

As mentioned before, a major challenge in the design of access schemes for MTC is the heterogeneity of the applications. Although current research trends now distinguish between at least two types of MTC (URLLC and mMTC), applications within each of these two categories still have heterogeneous reliability and latency requirements. Thus, the *second* research problem (Chapter 4) we tackle in this thesis is the analysis and design of RA schemes with SIC that can simultaneously satisfy different QoS guarantees (ref. [J3]).

1. We consider two different schemes where devices are grouped based on their QoS requirements. The first scheme is called the ACK-All scheme. In this scheme, MTC devices from all groups transmit simultaneously over the same radio resources in all stages of the transmission frame. The second scheme is called the ACK-Group scheme. In this scheme, MTC devices from different groups transmit in distinct stages. We present the case scenarios where our proposed RA schemes can service a larger system load.

2. In our schemes, successfully recovered packets are acknowledged via the feedback channel. A device that successfully receives an acknowledgement stops transmitting in the remaining part of the transmission frame. This behavior leads to dynamic and random reductions in the sizes of the groups during the same transmission frame. Based on this, we reformulate the AND-OR tree to accurately model these reductions and, consequently, accurately track the evolution of the error probabilities for each group in each iteration of the SIC process.

3. Finally, we use the derived expressions to design systems that can guarantee the QoS requirements of different groups with significantly high energy efficiency and high reliability. We also propose a guideline that allows us to design the access probabilities using the AND-OR tree, which was originally designed for asymptotically large systems, for a finite number of devices and resources.

1.2.3 The Analysis and Design of Grant-Free Non-Orthogonal Multiple Access

The random access schemes considered in Chapter 3 and Chapter 4 are orthogonal in nature. That is, the spectrum is divided into equal sized non-overlapping time-frequency units, and the number of users that can be supported is tightly coupled to the number of the available resource units. Nevertheless, in the presence of spectrum scarcity, Non-Orthogonal Multiple Access (NOMA) is more suitable for mMTC as it can support overloading, i.e., number of users supported is larger than the number of available resource units. Thus, the *third* and last research problem we tackle (Chapter 5) is the massive uncoordinated NOMA problem (ref. [J2]).

1. We consider an uplink grant-free NOMA setting where devices use pilot sequences as their signature. As these pilots are chosen uniformly at random, there is always a non-zero probability that two or more devices choose the same pilot sequence. In this case, a collision occurs and the receiver cannot distinguish the collided devices from one another. The receiver can only estimate their sum power. We propose to treat these signals as interference and prove that the interference power can be well approximated by a Poisson Point Process (PPP). This approximation proves useful in deriving closed form expressions of different performance metrics.

2. We first consider the case where all the devices transmit at the same fixed code-rate. We derive the expression of the outage probability for the case of joint decoding and successive interference cancellation. The evaluation of the exact expressions is shown to be daunting especially for the case of joint decoding. However, we show that by using the PPP approximation and the assumption of a massive number of devices, the expressions can be significantly simplified. The accuracy of the simplified expressions is demonstrated through simulations.

3. Then, we consider the case where all the devices transmit using rateless codes. In this case, the rate is determined on the fly and varies from slot to slot

based on the system load, received powers and interferers. The receiver stops transmissions by broadcasting a beacon when the throughput is maximized. We derive the expression for the maximum throughput for the case of joint decoding and successive interference cancellation. We also propose simplified expressions by using the PPP approximation and the assumption of a massive number of devices. The accuracy of the simplified expressions is demonstrated through simulations.

1.3 Thesis Outline

The rest of this thesis is organized as follows.

Chapter 2 starts with briefly introducing the concepts related to multiple access channels from an information-theoretic perspective. Then, it presents the literature review on uncoordinated and coordinated access. It also explains the motivation and challenges of the re-emerging non-orthogonal multiple access. The main contributions of this thesis can be found in Chapter 3-5. Each chapter has its one separate notations which are tabulated and presented at the beginning of every chapter for the readers' convenience. Chapter 3 focuses on SINR-based RA. In Chapter 4, we extend this to support multiple QoS requirements. In Chapter 5, we consider uncoordinated NOMA. Finally, a summary of this thesis and its major findings is provided in Chapter 6 along with some concluding remarks and future directions.

Chapter 2

Literature Review

In this chapter, we present the necessary background information for our research work. In Section 2.1, we briefly summarize the fundamentals of multiple access channels from an information-theoretic perspective. In Section 2.2, we explain the concept of random access and compare between some of the most popular protocols in this area. In Section 2.3, we explain how random access is applied to the control channels of some of the existing cellular networks, whereas data transmissions in cellular networks take place in a coordinated fashion. We also explain the challenges of supporting IoT with coordinated access, and we present some existing as well as potential future solutions. Finally, in Section 2.4, we explain the re-emerging concept of NOMA and its value to mMTC. We also shed light on the challenges hindering its realization.

2.1 Fundamentals of Multiple Access Channels

A multiple access channel is a channel shared by two or more transmitters that communicate with one receiver or more. Received signals are corrupted by noise and by mutual interference between transmitters. Each transmitter can have a sequence of messages to send that arrive at random instants of time. Thus, often, only a subset of these users have messages to transmit at the same time. Then, the problem is to determine how the time and frequency resources are to be shared among them.

In this thesis, we consider the simplest form of multiple access channels where all transmitters want to communicate with the same receiver.

In this section, we present the capacity limits of such multiple access channels. We start with the special case of a single transmitter. Then, we extend to a two-user case. The results herein can be easily extended to any number of transmitters. The channels considered here are all memoryless, with no channel side information and no feedback. Most importantly, we consider continuous Gaussian channels where the messages are corrupted by Additive White Gaussian Noise (AWGN). The extension of the work in this thesis to other channels, although possible, is considered out of scope.

2.1.1 Point-to-Point Channel

Figure 2.1 – Schematic Diagram of a point-to-point channel

A point-to-point channel consists of the communication between a single transmitter and a single receiver over a noisy channel (ref. Figure 2.1). The transmitter maps his k-letter message $\mathbf{b}$ into an n-letter codeword $\mathbf{x}$ drawn with some probability from the input alphabet $\mathcal{X}$. This mapping process is called encoding. The received signal $\mathbf{y}$ is a noisy version of $\mathbf{x}$ in the output alphabet $\mathcal{Y}$. The task of the receiver is to decode $\mathbf{y}$, which is defined as the process of obtaining an accurate estimate $\hat{\mathbf{x}}$ of the transmitted signal $\mathbf{x}$. We assume that the noise $\mathbf{w}$ is of the additive form such that

$$\mathbf{y} = \mathbf{x} + \mathbf{w}. \tag{2.1}$$

For AWGN, the channel transition probability of $\mathcal{X} \to \mathcal{Y}$ is expressed as

$$P_{Y|X}(\mathbf{y}|\mathbf{x}) = \left(\frac{1}{\sqrt{2\pi}}\right)^n e^{-\frac{||\mathbf{y}-\mathbf{x}||^2}{2}}. \tag{2.2}$$

A popular objective in information theory is to design a (n, M, ϵ_d) code such that one can communicate $M(= 2^k)$ different messages with the largest possible code rate

$$R_c := \frac{\log_2 M}{n}, \tag{2.3}$$

and a sufficiently small decoding error probability

$$\epsilon_d := \frac{1}{M} \sum_{i=1}^{M} P(\hat{\mathbf{x}} \neq \mathbf{x}_i), \tag{2.4}$$

assuming equally likely messages.

The efficiency of the code is characterized by R_c, and its reliability is characterized by ϵ_d. We will see that, in theory, ϵ_d can be made arbitrarily small by transmitting at higher powers. However, in practice, the transmit power is limited by the capabilities of the radio circuit power amplifiers. This power constraint can be expressed as follows:

$$\frac{1}{n}||\mathbf{x}||^2 \leq P_t. \tag{2.5}$$

Asymptotic Results for the Gaussian Point-to-Point Channel

For many years, the focus of information theory was on the achievable rates for asymptotically long codewords. Shannon [16] showed that a randomly selected code is good with high probability when its rate R_c is smaller than the channel capacity. The channel capacity is defined as the maximum number of information bits that can be received reliably in one channel use. He showed that the channel capacity of a Gaussian point-to-point channel, in the limit of $n \to \infty$, is

$$C(\text{SNR}) := \log_2(1 + \text{SNR}), \tag{2.6}$$

where the Signal-to-Noise Ratio (SNR) is defined as the ratio of the received signal power P to the noise power σ_w^2. For simplicity, and without loss of generality, we will take σ_w^2 to be one in all what follows. Lastly, Shannon showed that this capacity is

achievable with a Gaussian distributed input alphabet (as it will yield the optimal output distribution $\mathcal{CN}(0, 1 + P)$).

Non-Asymptotic Results for the Gaussian Point-to-Point Channel

A classical visualization of the communication channel is a pipe that allows the reliable transmission of C information bits per channel use. However, when n is not asymptotically large, the size of this pipe turns out to vary over time. In fact, these variations follow a Gaussian distribution with average C and variance V/n [17], where V is dubbed the channel dispersion parameter. Three main observations can be made from this analogy. The first one is that the achievable rate is a function of the block length. The second one is that the decoding error probability ϵ_d becomes more significant for short block transmissions. That is because the variations of the size of the pipe are random and unpredictable; therefore, the probability that the chosen code rate is larger than the size of the bit pipe increases as n decreases. The third observation, albeit intuitive, is that when $n \to \infty$, the effect of the channel dispersion is negligible and the size of the bit pipe, the channel capacity, is almost constant at all times as devised by Shannon.

The study of achievable rates in the finite block length regime started with Strassen [18] and was later significantly improved and extended by Polyanksiy, Poor and Verdu [19] as well as Hayashi [20]. The most recent results consist of a third-order approximation of the maximum achievable rate over a Gaussian point-to-point channel. This can be expressed as [21]

$$R_c(n, \epsilon_d) \leq C(P) - \sqrt{\frac{V(P)}{n}} + \frac{\log_2 n}{2n} + \mathcal{O}\left(\frac{1}{n}\right), \text{ where} \tag{2.7}$$

$$V(\mathrm{P}) := \frac{\log_2^2 e}{2}\left(1 - \frac{1}{(1+P)^2}\right), \text{ and} \tag{2.8}$$

$$Q(x) := \int_x^\infty \frac{1}{\sqrt{2\pi}} e^{-t^2/2} dt. \tag{2.9}$$

It is worthy of noting here that a Gaussian distributed input alphabet is sub-optimal.

In fact, Shannon stated this earlier on. The reasoning behind this is that a significant portion of the transmitted blocks will be transmitted at a power much less than the maximum allowed, thus, not efficiently utilized the available power resource.

Finally, another approach to the study of finite block length transmissions over the point-to-point Gaussian channel is based on error exponents. This research focuses on characterizing the rate of decay of the error probability as a function of n. Interested readers are referred to the works in [22, 23, 24, 25] for more information.

2.1.2 Multiple Access Channel

The information-theoretic approach for multiple access channels started in 1973 by Ahswede [26] and Liao [27]. Let us consider an example with two transmitters, U_1 and U_2, as shown in Figure 2.2. If these two users transmit over the same time-frequency resource unit, the received signal at the receiver can be expressed as

$$\mathbf{y} = \mathbf{x}_1 + \mathbf{x}_2 + \mathbf{w}, \tag{2.10}$$

where $\mathbf{x}_1$ and $\mathbf{x}_2$ are the transmitted codewords by U_1 and U_2, respectively. Moreover, $||\mathbf{x}_1||^2 \leq P_1$ and $||\mathbf{x}_2||^2 \leq P_2$, where P_1 and P_2 are the transmit power constraints of U_1 and U_2, respectively. As before, $\mathbf{w}$ is a vector of independent and identically distributed (i.i.d.) zero-mean AWGN with variance 1.

Asymptotic Results of the Gaussian Multiple Access Channel

A pair of rates $(R_{c,1}, R_{c,2})$ is said to be achievable for asymptotically long blocks if they belong to the capacity region of the multiple access channel. The set of achievable rates for a multi-user Gaussian Multiple Access Channel (MAC) is the set of all rates within the pentagonal capacity region devised by Cover [28] and Wyner [29]. This

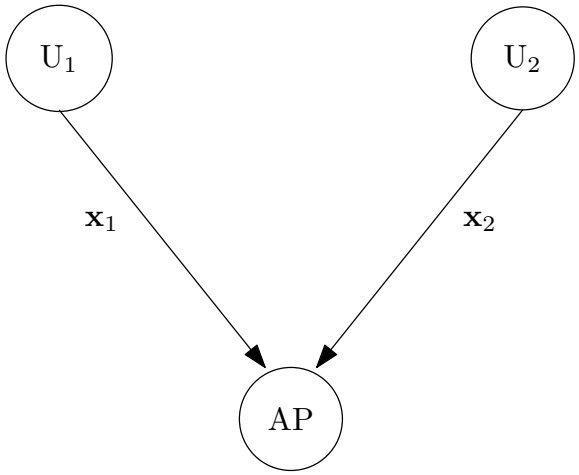

Figure 2.2 – Multiple Access Channel with Two Users and a Common Receiver

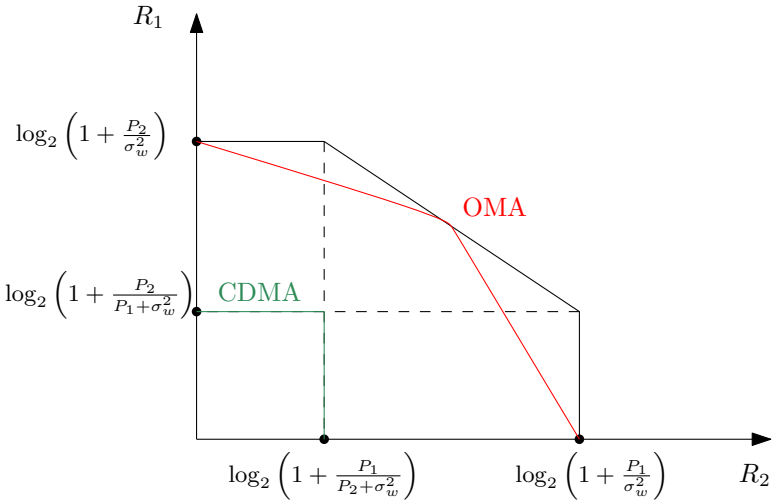

Figure 2.3 – Capacity Region of a Two User Gaussian Multiple Access Channel

capacity region is illustrated in Figure 2.3 and expressed below for the two-user case.

$$R_{c,1} \leq \log_2(1 + P_1) \tag{2.11}$$

$$R_{c,2} \leq \log_2(1 + P_2) \tag{2.12}$$

$$R_{c,1} + R_{c,2} \leq \log_2(1 + P_1 + P_2)\,. \tag{2.13}$$

The first two constraints dictate that users cannot exceed the capacity of their re-

spective point-to-point channel while the third constraint dictates that the sum rate $R_{c,1} + R_{c,2}$ cannot exceed the capacity of a point-to-point channel with the total power $P_1 + P_2$. Consequently, the two users cannot transmit at their maximum rates simultaneously.

In this context, we introduce three popular multiple access schemes and their respective capacity regions. The first one is Time Division Multiple Access (TDMA). Let us assume that each channel use is of duration T seconds. U_1 transmits in the first $T_1 = \alpha T$ seconds, where $0 \leq \alpha \leq 1$, and U_2 transmits in the remaining $T - T_1 = (1-\alpha)T$ seconds. In this way, the transmissions of the two users do not interfere with one another. As the users' transmissions are shorter in time, they can enhance their rates by scaling up their transmit power. The achievable rate region in this case is a subset of the aforementioned capacity region as shown below.

$$
\begin{aligned}
R_{c,1} &\leq \alpha \log_2\left(1 + \frac{1}{\alpha} P_1\right) \\
R_{c,2} &\leq (1-\alpha) \log_2\left(1 + \frac{1}{1-\alpha} P_2\right) \\
R_{c,1} + R_{c,2} &\leq \alpha \log_2\left(1 + \frac{1}{\alpha} P_1\right) + (1-\alpha) \log_2\left(1 + \frac{1}{1-\alpha} P_2\right).
\end{aligned}
$$

Now, let us assume that a channel use has a bandwidth W Hz. Frequency Division Multiple Access (FDMA) operates in a similar fashion except that both users transmit simultaneously for the entire T but over different frequencies [30]. To be specific, U_1 transmits over a fraction α of the bandwidth ($W_1 = \alpha W$), and U_2 transmits on the remaining portion $W - W_1 = (1-\alpha)W$. As the users are transmitting over narrower bandwidths, their transmissions are subject to less noise. In particular, the noise power is scaled down by a factor of α for U_1 and by a factor of $1-\alpha$ for U_2. Thus, the achievable rate region is the same as that of TDMA.

In Code Division Multiple Access (CDMA), each user uses a long spreading code to spread its message over the entire duration T and the entire bandwidth W, simultaneously with other users. The receiver receives the superposition of all the transmitted messages. Each user is decoded by considering the other users' signals as interference.

Thus, we have

$$R_{c,1} \leq \log_2\left(1 + \frac{P_1}{1+P_2}\right), \text{ and}$$
$$R_{c,2} \leq \log_2\left(1 + \frac{P_2}{1+P_1}\right).$$

The capacity regions of these schemes are illustrated in Figure 2.3 for the two user case. Although Equation 2.13 indicates that both users cannot transmit at their maximum rate simultaneously, the corner points of the capacity region indicate that one user can in fact achieve its point-to-point capacity while the other user's rate is non-zero. These corner points are achievable via simultaneous transmissions of both users over the entire time and bandwidth. The receiver decodes the superposed signals in two stages. In the first stage, the receiver decodes the strongest signal while regarding the second one as interference. Then, if successful, it uses SIC to cancel the recovered signal from $\mathbf{y}$. In the second stage, it will decode the weaker signal which is no longer suffering from any interference.

Interestingly, Orthogonal Multiple Access (OMA), e.g., TDMA and FDMA, can achieve one optimal point when $\alpha = \frac{P_1}{P_1+P_2}$. That is, it is optimal when the resources are split in proportion to their received powers. However, this resource allocation scheme is unfair as it allocates more resources to the user with the better channel. The remaining points can be achieved through time sharing or rate splitting.

Non-Asymptotic Results for the Gaussian Multiple Access Channel

Upon first glance, the achievable capacity region for multiple access channels in the finite block length regime might seem readily derivable from Equation 2.11, Equation 2.12, and Equation 2.13 to the following:

$$R_{c,1} \leq C(\mathrm{SNR}_1) - \sqrt{\frac{V(\mathrm{SNR}_1)}{n}} + \frac{\log_2 n}{2n} + \mathcal{O}\left(\frac{1}{n}\right), \tag{2.14}$$

$$R_{c,2} \leq C(\mathrm{SNR}_2) - \sqrt{\frac{V(\mathrm{SNR}_2)}{n}} + \frac{\log_2 n}{2n} + \mathcal{O}\left(\frac{1}{n}\right), \tag{2.15}$$

$$R_{c,1} + R_{c,2} \leq C(\mathrm{SNR}_1 + \mathrm{SNR}_2) - \sqrt{\frac{V(\mathrm{SNR}_1 + \mathrm{SNR}_2)}{n}} + \frac{\log_2 n}{2n} + \mathcal{O}\left(\frac{1}{n}\right). \tag{2.16}$$

However, the expressions on the right-hand side of the above inequalities are only upper bounds and are most likely not achievable. This is because they assume that $||\mathbf{x}_1 + \mathbf{x}_2||^2 = ||\mathbf{x}_1||^2 + ||\mathbf{x}_2||^2$, where, in reality, the left-hand side is strictly less.

The second-order approximation of the capacity region of a multi-user Gaussian MAC was derived in [31] and was shown to be quite cumbersome as it involves the evaluation of multi-dimensional Gaussian distribution. This capacity region is given below for interested readers for the case of two users.

$$\begin{pmatrix} R_{c,1} \\ R_{c,2} \\ R_{c,1} + R_{c,2} \end{pmatrix} \in \begin{pmatrix} C\left(P_1\right) \\ C\left(P_2\right) \\ C\left(P_1 + P_2\right) \end{pmatrix} - \frac{1}{\sqrt{n}} Q^{-1}\left(\epsilon_d, \mathbf{V}\left(P_1, P_2\right)\right) + \mathcal{O}\left(\frac{1}{n}\right)\mathbf{1}, \tag{2.17}$$

where

$$Q^{-1}\left(\epsilon_d, \mathbf{V}\right) = \left\{\boldsymbol{\alpha} \in \mathbb{R}^3 : \Pr\left(\mathcal{N}(\mathbf{0}, \mathbf{V}) \leq \boldsymbol{\alpha}\right) \geq 1 - \epsilon_d\right\}, \tag{2.18}$$

and the channel dispersion matrix is

$$\mathbf{V}\left(P_1, P_2\right) := \begin{pmatrix} V(P_1) & V_{1,2}(P_1, P_2) & V_{1,3}(P_1, P_2) \\ V_{1,2}(P_1, P_2) & V(P_2) & V_{2,3}(P_1, P_2) \\ V_{1,3}(P_1, P_3) & V_{2,3}(P_1, P_2) & V(P_1 + P_2) + V_3(P_1, P_2). \end{pmatrix} \tag{2.19}$$

For OMA, we know that the users' transmissions do not overlap. Thus, the probability of decoding $\mathbf{x}_1$ is independent of the probability of decoding $\mathbf{x}_2$, and vice versa. Then, ϵ_d can be expressed as $\beta_1\epsilon_d + \beta_2\epsilon_d - \beta_1\beta_2\epsilon_d^2$, where $\beta_1\epsilon_d$ and $\beta_2\epsilon_d$ are the probabilities of decoding $\mathbf{x}_1$ and $\mathbf{x}_2$ erroneously, respectively. For $0 \leq \alpha, \beta_1, \beta_2 \leq 1$, the achievable

set of rates for OMA are bounded by the following constraints:

$$R_{c,1} \leq \alpha C\left(\frac{P_1}{\alpha}\right) - \sqrt{\frac{\alpha}{n}} V\left(\frac{P_1}{\alpha}\right) Q^{-1}(\beta_1 \epsilon) + \mathcal{O}\left(\frac{1}{n}\right), \tag{2.20}$$

$$R_{c,2} \leq (1-\alpha) C\left(\frac{P_2}{1-\alpha}\right) - \sqrt{\frac{1-\alpha}{n}} V\left(\frac{P_2}{1-\alpha}\right) Q^{-1}(\beta_2 \epsilon) + \mathcal{O}\left(\frac{1}{n}\right). \tag{2.21}$$

The main take-away from these results is that OMA schemes are sub-optimal for short block transmissions. In fact, it was shown in [32] that the gap to the capacity increases with the increase in the number of users. These results emphasize the compatibility of NOMA schemes with the requirements of MTC. We will see next that current cellular networks are unfortunately based on OMA and thus not suitable to provide IoT and M2M services.

2.2 ALOHA-based Random Access Protocols

2.2.1 Brief History of ALOHA

Abramson proposed the ALOHA random access protocol at the University of Hawaii in 1970 [33] to provide a wireless connection between the computer resources of different islands of the state of Hawaii. At the time, only wired-based communications, e.g., dial-up telephone, could allow remote access to large information processing systems. Moreover, centralized control was used where a central control system determined distinct channel access times to the transmitters to avoid interference. In general, centralized control is complex to design and incurs large overhead especially in the case of brief and infrequent transmissions. To this end, ALOHA is a wireless communication protocol that multiplexes a large number of transmitters with bursty traffic over one radio channel. In ALOHA, transmitters access the channel at random, yet independent, times and in an asynchronous fashion. Thus, it does not require any centralized control and incurs almost no overhead.

In ALOHA, a transmitted packet is said to be successful if it is received without interference, and if it is decoded successfully. Interference is caused by other packets

transmitted within the duration of the packet. Packets suffering from interference are said to have collided, and the data is said to be lost. On the other hand, decoding errors depend on random noise errors introduced by the channel itself. Erroneously decoded packets can be detected via a Cyclic Redundancy Check (CRC). In case of no decoding error, transmitters are sent acknowledgements via a feedback channel. If a transmitter does not receive an acknowledgement within a certain time frame, it re-transmits a copy of its packet at another randomly chosen instant of time (ref. Figure 2.4).

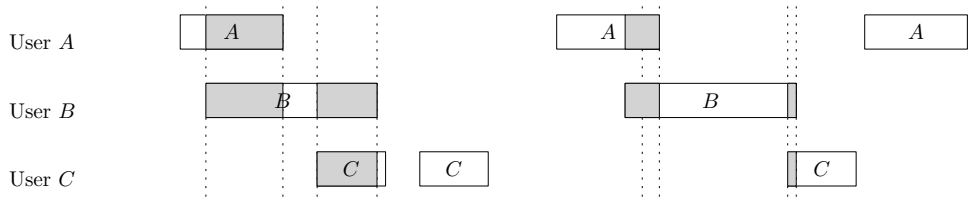

Figure 2.4 – Schematic Diagram of ALOHA with 4 transmitters. The shaded areas correspond to interference between overlapping segments of different packets. Packets suffering from interference are said to have collided and irretrievable.

Now, let us model the inter-arrival times of the new packets as well as the re-transmission times of previously unsuccessful packets as independent and exponentially distributed processes. For a small number of transmitters with low duty cycles, the probability of a packet overlapping and, thus, interfering/colliding with another packet is small. More specifically, for a packet arrival rate of λ packets per second, where each packet is τ seconds long, the normalized throughput is equal to $\lambda\tau e^{-2\lambda\tau}$. Moreover, the maximum normalized throughput is $\frac{1}{2}e^{-1} \sim 0.184$.

A few years later, Slotted ALOHA (S-ALOHA) was proposed by Roberts [34] and Abramson in 1977 [35]. S-ALOHA pre-defines a set of contiguous time slots of equal duration. The transmitters should align the start of their packets to the start of any of the slots (ref. Figure 2.5). By eliminating asynchronous transmissions, the maximum achievable throughput is doubled ($e^{-1} \sim 0.368$). However, it is worthy of noting that this improvement is based on the assumption of equal packet lengths. Otherwise, the loss in throughput due to the wasted portions of the time slots can be

larger than the factor of two improvement. Furthermore, S-ALOHA requires perfect synchronization which requires more overhead than ALOHA. Motivated by this and by the fact that long packets are more susceptible to interference, variants of the protocol were proposed to reserve resources a priori for long packets, e.g., Reserved ALOHA (R-ALOHA) [36] and Demand Assigned Multiple Access (DAMA).

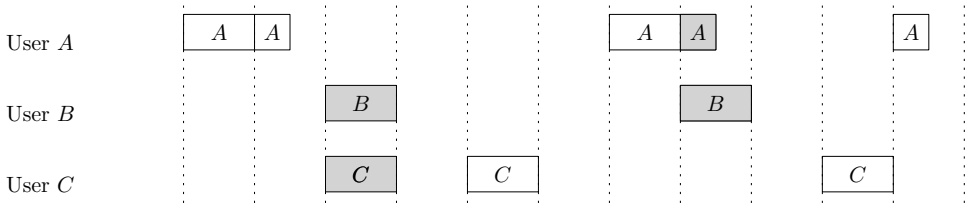

Figure 2.5 – Schematic Diagram of ALOHA with 4 transmitters

Collision avoidance protocols were later proposed to minimize the probability of collision. To this end, Carrier Sense Multiple Access (CSMA) was introduced to avoid collisions at the expense of hardware and software complexities. In CSMA, prior to transmission, the transmitter senses the environment to see if the channel is occupied, i.e., any other transmission is already taking place. Sensing can be done by measuring some received power and comparing it to a predetermined threshold. Furthermore, Carrier Sense Multiple Access with Collision Avoidance (CSMA-CA) was proposed where the back-off follows an exponential distribution so as to minimize the probability of a second collision. The intuition behind this is that the more collisions occur, the more congested the network is. Thus, retransmissions need to be less frequent (more spaced apart in time). However, CSMA suffers from the hidden terminal problem, i.e., when the transmitters are not within the sensing range of one another. CSMA is also unsuitable for long range communications where the propagation delay is larger than the packet transmission time, e.g., satellite communications and broadband hybrid fiber coaxial networks.

In the above ALOHA-based protocols, slots can be categorized into three types: idle slots (no packets received), singleton slots (one packet received) and collision slots (multiple packets received). These protocols assume that only interference-free

copies of the packet can be correctly recovered, i.e., singleton slots. Thus, idle slots are wasteful of resources. Moreover, collision slots are wasteful of both resources and power. Next, we will explore collision resolution protocols that can retrieve some of the collided packets and, thus, achieve a higher throughput.

2.2.2 ALOHA and Successive Interference Cancellation

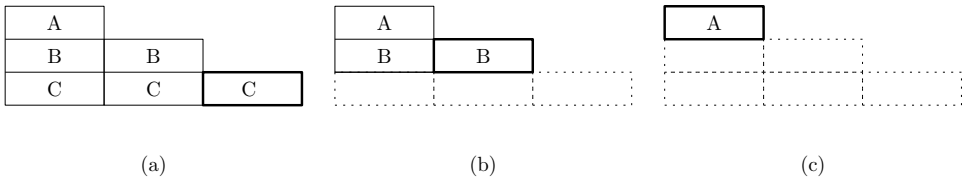

Figure 2.6 – An example of Random Access with Successive Interference Cancellation

Amongst the numerous ALOHA variants that were proposed, the major break-through started with the introduction of SIC for contention resolution. A simple example is illustrated in Figure 2.6 where three types of packets are transmitted over three time slots. Only the third slot is an idle slot. Thus, packet C can be recovered at the receiver. Now, let us suppose, that each packet header contains pointers to all other slots containing the same copy. The receiver could exploit this information to cancel packet C from slot 1 and slot 2. Assuming packet C can be perfectly cancelled from slot 2, a new idle slot is created containing packet B as shown in Figure 2.6b. Similarly, packets B and C can be cancelled from slot 1 allowing the recovery of packet A as shown in Figure 2.6c. For this particular example, the throughput is improved by factor of 3 with SIC.

One of earlier protocols using SIC is Contention Resolution Diversity Slotted Aloha (CRDSA) [37]. In CRDSA, users transmit two copies of their information. It is successful if one of the two copies is received with no interference. The recovered packet is assumed to have a pointer indicating the location of its replica. Then, the Access Point (AP) uses SIC to cancel out the interference of this packet in the other slot. The interference cancellation process is iterated until most of the packets in the frame

that were lost due to collision are recovered. This leads to the potential recovery of more packets and eventually a higher throughput. Packets are also assumed to be appended by orthogonal preambles that enable the receiver to estimate the channel parameters such as carrier frequency, amplitude and timing estimation which is identical for all copies. This estimation is possible provided that these parameters remain constant for the duration of the slot. However, the phase would be different in different slots.

Other random access protocols using SIC are SICTA [38], Zigzag [39] and Sigsag [40]. The latter two were proposed especially for asynchronous transmissions. In general, for many years, these protocols were shown to exhibit excellent delay-throughput characteristics when traffic is low. On the other hand, coordinated multiple access schemes, such as those based on TDMA, FDMA and CDMA, were seen suitable for steady and relatively heavy traffic. Nevertheless, we saw in Section 2.1 that orthogonal schemes are strictly sub-optimal in the finite block length regime and, thus, strictly sub-optimal for the transmission of the short payloads of mMTC. Similarly, we will later see that the above statement does not hold for mMTC. In particular, we will show that coordinated access is sub-optimal in massive access mainly due to the large overhead that scales larger than the packet lengths.

2.2.3 Coded Random Access

Although the aforementioned works on random access and SIC demonstrated a promising increase in the system throughput, the analysis and design of these protocol was still quite complex. It was not until the seminal paper [41], that a solid analytical framework was available for analysis and optimization. In Irregular Repeat Slotted ALOHA (IRSA) [41], the number of transmissions for every user is a random variable whose distribution is predetermined by the Base Station (BS)[1] and broadcasted at the beginning of each transmission period. Liva [41] drew a remarkable analogy between

[1]We will use the term Base Station (BS) and Access Point (AP) interchangeably in this thesis. In practice, BS is usually used in cellular networks whereas AP is usually used in WiFi. Both provide users with wireless access to wired networks.

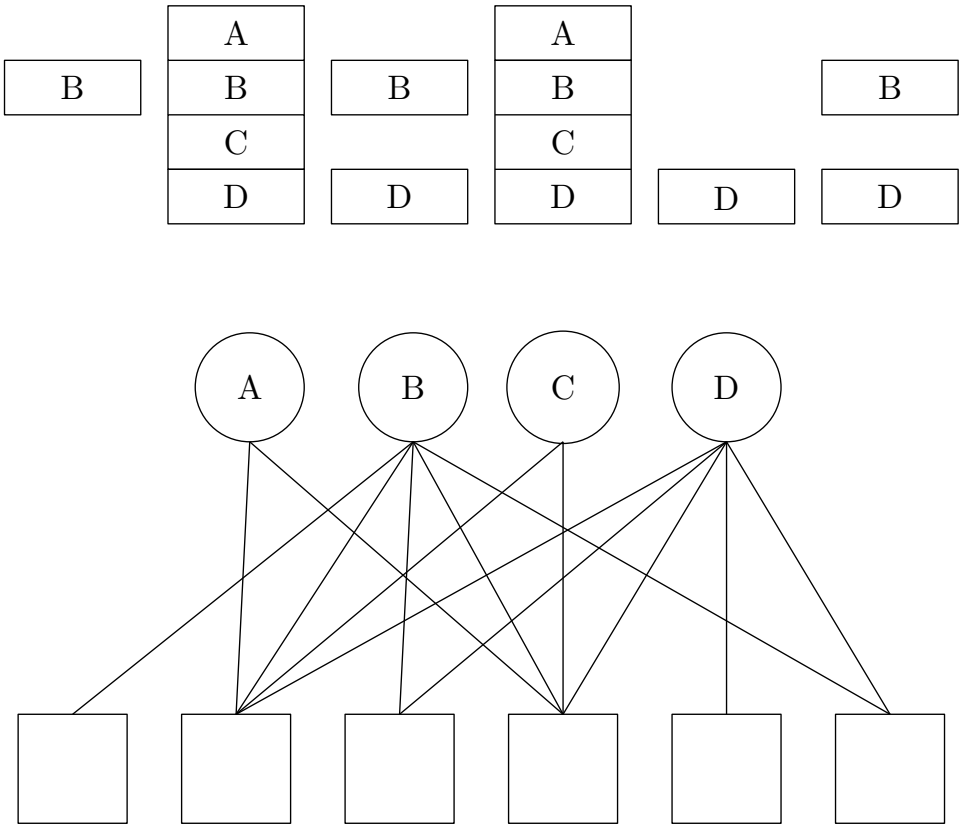

Figure 2.7 – An example of Irregular Repeat Slotted ALOHA with 4 Transmitters and 6 Slots. The circles represent the users and the squares represent the slots.

RA and SIC and iterative message passing decoding codes-on-graph for the BEC, i.e., Low Density Parity Check (LDPC) [42], Luby Transform (LT) [43], Raptor codes [44], etc. Similar to codes-on-graph, IRSA can be represented in terms of a bipartite graph with two sets of nodes representing the number of users and the number of slots. The nodes are connected by edges representing the packet transmissions (ref. Figure 2.8). Moreover, for asymptotically large sets of nodes, the graph can be approximated by a tree topology, for which the AND-OR tree analysis [45] can be used to analyze the convergence of the SIC process and to calculate the average recovery error probability. It can also be used to find the necessary probabilities and the maximum load that can be supported.

Many extensions have been proposed since that have further emphasized the scalability of RA with SIC and its capability of supporting a large number of devices

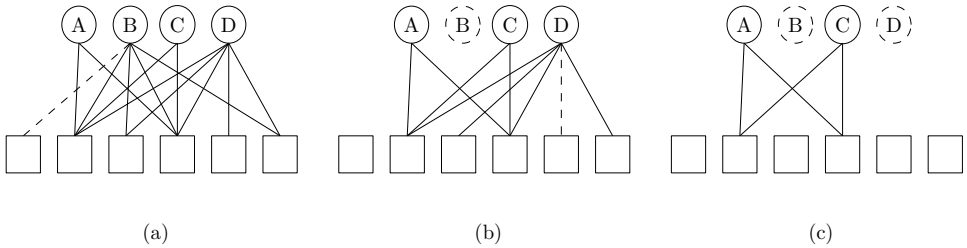

Figure 2.8 – A Graphical Representation of SIC in Irregular Repeat Slotted ALOHA with 4 Transmitters and 6 Slots

without prior identification [46, 47, 48]. By requiring little signalling overhead and centralized processing, it is being considered as a possible candidate for future M2M communications [49].

2.3 The Random Access Channel in LTE

From Chapter 1, we learnt that a general consensus has been reached on the need for the fifth generation of communication systems to support new services, besides the traditional voice and data services, namely, IoT and M2M services. In Long Term Evolution Advanced (LTE-A), devices are required to establish an air interface connection prior to data transmission. Access requests are transmitted in an uncoordinated manner over the Random Access Channel (RACH) using protocols similar to those explained in Section 2.2. The RACH is a sequence of time-frequency resources called RA slots. It constitutes the transport layer channel and is responsible for the management of the channel access requests triggered by the end devices.

However, the emerging IoT and M2M services are fundamentally different in nature and requirements from the conventional Human to Human (H2H) communications. H2H communications are characterized by the transmissions of large amounts of data, with high data rates, and a somewhat predictable traffic. On the other hand, M2M communications are characterized by the transmissions of very small amounts of data, with low data rates, low power budget, low computational and processing capabilities and a sporadic traffic. This calls for major changes in the air interface and communi-

cation protocols. To this end, the enhancement of the operation of the random access channel of Long Term Evolution - 4G (LTE) and LTE-A has been identified as one of the key challenges for future MTC [50]. We will explain this operation next. In what follows, we refer to a user as User Equipment (UE) and to the BS as Evolved Node-B (eNB) as per LTE terminology.

2.3.1 Random Access Procedure

In LTE-A, RA can be used for initial access to establish a radio link, to perform uplink synchronization, to schedule a request, to re-establish a radio link after failure and to handover from one eNB to another. The minimum resource scheduling unit of downlink and uplink transmission is referred to as a Resource Block (RB). In the frequency domain, a RB consists of 12 sub-carriers each with a bandwidth of 180 kHz. In the time domain, a RB consists of one sub-frame each with a duration of 1 ms. Random access takes place over allowable time slots, called an Access Grant Time Interval (AGTI) or RA opportunity. The time-frequency resource on which RA is performed is called the Physical Random Access Channel (PRACH). The RA slots have a bandwidth of 1.08 MHz which corresponds to six RBs, and a basic duration of 1 ms.

The available set of PRACH resources for transmitting preambles is broadcasted by the eNB. The RA procedure, often referred to as the four-way handshake, is summarized as follows:

Message 1: Preamble transmission

In the first available RA slot, the UE randomly chooses an orthogonal preamble, an Orthogonal Frequency Division Multiplexing (OFDM)-based signal, uniformly at random from a predetermined set. The selected preamble acts as the UE's digital signature. The periodicity of the RA slots and the number of preambles are determined by the eNB. In general, each LTE cell has 64 preambles created by the Zadoff-Chu sequence. However, some of these preambles are reserved for contention-free RA.

Message 2: RAR over PDSCH

The eNB detects the preambles by performing cross-correlations between the received signals and all the preambles from the set. When a preamble is auto-correlated with itself, it results in a Channel Impulse Response (CIR). Furthermore, due to the orthogonality nature, when a preamble is cross-correlated from another preamble from its set, it will result in the all-zero sequence. The eNB will decode the transmitted access requests to estimate the channel parameters, e.g., timing off-set, channel gains, etc. If two or more UEs choose the same preamble, the received signals appear as a single transmission going through multiple fading paths which allows the eNB to detect the collisions. Collisions are discarded. The eNB sends the successful UEs an Random Access Response (RAR) with an ID indicating the time-frequency slot in which the preamble was sent, uplink scheduling grant, respective Timing Advance (TA), assigned Physical Uplink Shared Channel (PUSCH) resources, Cell Radio Network Temporary Identifier (C-RNTI) and optional back-off offset. The latter is necessary in case of failure to reduce the probability of another collision by dispersing the access attempts over time. We will elaborate on this later.

Message 3: Connection Request on PUSCH

The RAR contains different sub-headers dedicated to the successfully detected preamble. If the RAR does not contain the UE's preamble, it flags access attempt as a failed attempt and re-transmits based on the indicated back-off time. Similarly, if a UE does not received a RAR within a specified time interval, it also flags its access attempt as a failed attempt. The rest of the devices transmit a Connect Request message to the eNB in the granted resources in Message 2. This takes place using Hybrid Automatic Repeat Request (HARQ). It is worthy of noting that when two or more UEs involved in a collision are at the same distance from the eNB, the eNB is very likely not to be able to detect a collision as their received signals will be received constructively. In this case, a collision will occur by their Message 3 in every transmission attempt until the maximum number of retransmissions is reached. Then, the involved UEs

will declare a failed access attempt.

Message 4: Contention Resolution

Finally, the eNB answers the received Connection Requests with a Contention Resolution message. The UEs that do not receive a Contention Resolution message will re-attempt access in the next available RA after any necessary back-off. After a predetermined number of failed access attempts, the network is declared as unavailable to the UE and the issue is raised to the upper layers.

The RA procedure is either contention-based or contention-free. In the contention-free case, devices use dedicated preambles for access, leading to lower access latencies in comparison to the contention-based approach. It is mainly used in scenarios with critical delay requirements, e.g., a handover. In most cases, 10 out of the 64 available preambles are reserved for contention-free access.

Once an MTC device has been granted access, it is scheduled to specific radio resources over which data transmission takes place in a deterministic manner. As mentioned above, there are approximately 54 preambles available for contention-based access. Moreover, access opportunities have a period of 5 ms which is equivalent to a maximum system capacity of 10,800 preambles per second [51]. This capacity is achievable only in the absence of collisions, i.e., the case where each preamble is chosen by only one device. Thus, in reality, the number of successful access attempts is less than 10,800 per second. In fact, the number of successful access attempts is inversely proportional to the network load, as an increased load leads to an increased number of collisions.

Therefore, it is straightforward to see that standard LTE will not scale well with the massive access attempts, leading to a sharp degradation of the quality offered to M2M as well as H2H services. This performance degradation is in the form of long and unpredictable access delays, larger access failure rates, larger energy expenditure due to the increased number of retransmissions, and resource wastage.

2.3.2 Random Access Channel Congestion Control

We now review some of the solutions that have been proposed and even standardized to mitigate the RACH overload problem.

Access Barring

The basic idea is that the eNB broadcasts a certain access probability (usually a function of the system load). Prior to every access attempt, the UE draws a random number between zero and one. If the selected number is larger than the access probability, the device is blocked and not allowed access. Otherwise, the device follows the procedure outlined above. Although Access Class Barring (ACB) is efficient in reducing collisions, it does not suffice as a stand-alone solution as it incurs longer access delays.

Back-Off Schemes

As mentioned above, UEs with failed access attempts can refrain from re-attempting access in the next available RA opportunity. By delaying access through random back-off times, the access attempts are dispersed over time and the number of collisions is reduced. Naturally, the random back-off time grows with the network load.

Dynamic Resource Allocation

LTE defines 64 different configurations for the RACH. These configurations represent different tradeoffs between the amount of RA opportunities and the number of data channels (resources available for data transmission). Increasing the number of RA slots implies decreasing the amount of resources available for data transmission. With dynamic resource allocation, these resources are dynamically based on the changes in the network load. Some self-optimizing algorithms have been proposed in this context [52, 53, 54].

Pull-Based Schemes

In pull-based schemes, the eNB controls the number of devices that access the network by paging them individually. If the UE receives a paging message and requires access, it will follow the procedure outlined above. Otherwise, it will ignore the message. However, for large networks, this scheme incurs far too much overhead and is, thus, inefficient.

Clustering

Devices can be grouped into clusters based on their application type, QoS requirements, geographical location and Channel State Information (CSI). Each cluster is assigned a cluster head which is responsible for communicating with the eNB on behalf of the cluster members. This can be enabled with peer-to-peer or Device to Device (D2D) communication technologies. By restricting communication to the eNB to the cluster heads, the number of access requests is reduced. Consequently, the access failure rate is reduced as well.

Collision Detection and Resolution

Other schemes have aimed at finding different ways to detect and resolve collisions. For example, users can send their IDs during the four-way handshake. If only one ID is sent, the BS can recover this. In case of a collision, eNB receives the superposition of two or more IDs and is unable to retrieve any of them. Failure to recover a valid ID allows the eNB to detect a collision. By increasing the collision detection success rate, fewer data channels will be wasted by being allocated to multiple users simultaneously. Alternatively, if the users who chose the same preamble sequence are at different distances from the receiver, their preambles will be received with distinguishable multi-path propagation delays allowing the eNB to detect a collision. In this case, the eNB can either choose to serve the user with the smallest propagation delay or, if practical, can choose to serve them all. For the latter approach, users are

said to be identified via their different propagation delays. Note that the feasibility of the latter approach lies in having static users such that their timing advance does not change significantly over time [55].

2.3.3 Other Challenges

Aside from the overload problem of the RACH, we now present some of the other challenges related to providing M2M services over cellular networks.

Energy Efficiency

Unlike human-type devices that can be plugged into charging ports whenever the battery is running low, many machine-type devices will be deployed in remote areas and will be required to operate without a battery replacement for more than 10 years. Thus, enabling technologies and algorithms should consist of low computational, processing and storage requirements as well as low duty cycles. For instance, protocols involving many retransmissions are desirable here.

QoS Requirements: Latency and Reliability

One of the major challenges in providing M2M services is addressing the diverse QoS requirements for the heterogeneous applications. These QoS requirements are most commonly defined by the latency and reliability requirements, with the latter encompassing cases of failed attempts as well as network unavailability. LTE has defined a total of 16 different classes. In the case of network overload, the eNB invokes ACB where it transmits a set of probability factors and barring timers as part of the system information corresponding to the set of different classes available. For Extended Access Barring (EAB) [56], devices that belong to lower classes, i.e., delay-tolerant applications, are completely barred from accessing the network.

Effect on H2H Communications

Most solutions provide some form of separation between H2H and M2M. That is, the preambles are divided into two subsets allocated exclusively to H2H and M2M communications, respectively. Otherwise, the performance of M2M can highly jeopardize that of H2H.

Control Overhead

In coordinated access, the control overhead will scale larger than the payloads themselves, which is highly inefficient. Thus, the four way handshake in LTE needs to be revised and re-optimized.

2.4 Non-Orthogonal Multiple Access

2.4.1 The Re-emergence of NOMA

The term NOMA was previously known as superposition coding and was first proposed by Cover in 2011 [57]. In fact, NOMA has been used in previous wireless systems, such as a 3G Wide-band Code Division Multiple Access (WCDMA) system (ref. Section 2.1. Higher throughput performances are achieved in 3.9/4G by adopting Orthogonal Frequency Division Multiple Access (OFDMA) and Single Carrier Frequency Division Multiple Access (SCFDMA) as the multiple access schemes. The OFDM signal is robust against multi-path interference and is in harmony with Multiple Input Multiple Output (MIMO) technologies.

Nonetheless, the re-emerging NOMA exploits a new domain that has been underutilized so far: the power domain. In NOMA, the multiplexing of signals with disparate power levels allows for the efficient use of SIC and point-to-point capacity achieving forward error correction codes, e.g., Turbo code and LDPC. NOMA can score significant performance gains over OFDMA in many case studies when the power levels are sufficiently disparate.

Conventional resource allocation strategies, such as water filling strategies, allocate more resources, i.e., power, time and bandwidth, to users with strong channel conditions. These strategies are based on the assumption that the channel conditions of the different users follow the same distribution. In other words, users are equally likely to have good channels and, thus, are allocated an equal number of times and an equal number of resources, on average. However, this is obviously not true in the case of static devices that are to operate from the same location for over 10 years. In this context, NOMA has re-emerged to emphasize the importance of allocating more power to the far users such that they have acceptable QoS. In particular, the far users should be allocated enough power to be able to reliably recover their own messages directly by treating the other users' information as noise. On the other hand, the near user needs to first recover the message of the far user. Then, it subtracts this message from the received signal to recover to its own message. Thus, the feature of NOMA is not only to yield large throughput (larger than that of OMA) but also to ensure fairness.

For the uplink transmission, this concept of allocating more power to the far users cannot be applied as it dictates that the far users expend more power in comparison to the near users, which is unfair. However, uplink NOMA is still desirable for future MTC for many reasons: it can serve more users than OMA over the same resources due to its capability for overloading, the processing complexity due to the superposition of the signals is at the BS which suits the characteristics of MTC, and, finally, it is capacity achieving for short packet transmissions (ref. Section 2.1). All in all, NOMA is forecasted to play an essential role in the future generation of cellular networks, for both uplink and downlink transmissions.

2.4.2 Challenges of NOMA

In this section, we highlight the main challenges hindering the progress and feasibility of both uplink and downlink NOMA.

Processors

Advanced detection and decoding techniques are essential to separate and recover the individual signals that are received in superposition. Moreover, the receivers require high and fast processing and storage capabilities to keep the end-to-end delay low.

Power Constraints

In the uplink, the users' maximum transmit power is dependent on their battery power. Users can independently transmit with their maximum battery power. However, the performance of NOMA (especially with SIC) is very sensitive to the differences between the channels gains of these users in this case. Thus, centralized power control is necessary when the channel gains are too close to maintain these differences and, consequently, maintain good performance.

SIC Error Propagation

When the signals are not recovered correctly at a certain stage of SIC, interference cannot be cancelled perfectly. This will lead to the propagation errors whose impact increases with the increase in the number of stages. This significantly degrades the performance of NOMA in reality. Some potential solutions include non-linear detection techniques that can marginalize this effect [58].

Residual Time-Offset

For downlink transmissions, the BS controls the transmissions and thus can easily achieve near perfect synchronization. However, the case is not as simple in the uplink due to the spatial distribution and mobile characteristics of the users. In asynchronous communications, the superposed OFDM symbols are not aligned well in time, and this relative timing offset has a significant impact on the performance of NOMA users [63]. Accurate information on multiple symbols is required to be available at the receiver for reliable user detection and interference cancellation.

Extension to Multi-cell

In multi-cell scenarios, the interference is both of intra-cell and inter-cell nature. Combatting interference in a multi-cell scenario requires joint pre-coding across neighboring cells. The feasibility of such joint designs is in question [59, 60, 61].

User and Resource Allocation

In current LTE, the UEs perform power control to have the same receive power. The power control process is a coordinated and centralized process. In NOMA, UEs have to perform power control to have different receive powers. This also requires coordination. Let us consider a simple two-user scenario: one near and one far. For a required QoS guarantee, the throughput of the near user is dependent on the distance of the far user to the receiver. More specifically, for a target throughput R^*, the far user needs to be located beyond a distance D from the receiver such that the following condition is satisfied.

$$R^* = \left(\frac{1 + P_1 D^{-\alpha}}{P_2} \right)^{-\frac{1}{\alpha}} \tag{2.22}$$

This lays the selection criteria necessary for user pairing. The extension of this concept to a multiple user scenario is very complex. In fact, the optimal resource and user allocation problem is NP hard. Alternatively, the development of powerful yet practical resource and user allocation is a well-investigated topic [62, 63, 64, 65, 66, 67, 68, 69, 70, 71, 72]. In brief, it is forecasted that future communications will involve hybrid multiple access schemes of OMA and NOMA. We will elaborate on this later.

2.4.3 NOMA for Uplink Transmissions: Code Domain

As mentioned before, power control for uplink transmissions is needed for NOMA to perform well. However, NOMA can also multiplex users in the code domain, e.g., CDMA. We list some of the recent schemes proposed in this direction.

Low Density Spreading

In Low-Density Spreading (LDS) [73], the low density property is based on each user transmitting in a small fraction of the available resources units, e.g., frequency channels. However, there can be more than one user occupying the same resource unit. In LDS-CDMA, low-density spreading sequences are used to restrict the effect of interference on each chip of CDMA systems. On the other hand, LDS-OFDM can be thought of as a combination of LDS-CDMA and OFDM. In LDS-OFDM, after the information symbols are spread across the low-density spreading sequences, the chips are transmitted on a set of orthogonal frequency channels [73].

Sparse Code Multiple Access

Sparse Code Multiple Access (SCMA) is based on LDS-CDMA. It combines bit mapping and bit spreading [74, 75, 76]. SCMA maps the information bits to sparse codewords drawn from a multi-dimensional code book. Then, SCMA multiplexes the codewords over a set of orthogonal time-frequency resources, e.g., OFDM. The sparsity of the code allows the receiver to use low-complexity message passing algorithms to recover the information.

Pattern Division Multiple Access

In MIMO channels, the diversity order of a layer increases with its detection order. Inspired by this property and the fact that a multiple access channel resembles a virtual MIMO channel, Pattern Division Multiple Access (PDMA) maps the data to a resource group with a specific diversity order. The diversity order of a resource group is a function of the number of time, frequency and spatial resources it consists of [77]. The disparate diversity orders aim to mitigate propagation errors in the SIC process.

Spatial Division Multiple Access (SDMA)

Spatial Division Multiple Access (SDMA) [78, 79, 80, 81] assigns each user a unique CIR which allows the receiver to detect each user separately and accurately. SDMA is of particular benefit when the number of transmitting users is moderate yet larger than the number of receive antennas. However, it is suspected that applying SDMA to networks with a large number of users is challenging due to the degradation in the accuracy of CIR estimation.

Uplink MUSA

The uplink Multiple User Spectrum Access (MUSA) scheme is based on the enhanced Multi-Carrier Code Division Multiple Access (MC-CDMA) scheme. MUSA can score large performance gains due to its advanced low correlation spreading sequences, SIC techniques at the receiver, and linear processing. MUSA is of particular benefit when the ratio of the number of users to the number of resources (aka the overloading factor) is relatively high. Nonlinear detection algorithms such as SIC, Maximum a Posteriori (MAP), or Maximum Likelihood (ML) are used at the receiver side to separate the multiplexed users.

2.4.4 Software Defined Multiple Access

As mentioned before, user and resource allocation, whether static or dynamic, is very challenging in NOMA systems. In fact, the optimization problem is of combinatorial nature and is computationally prohibitive. Thus, new low-complexity algorithms are needed to determine the optimal user clustering. On the other hand, authors can restrict cluster sizes to pairs and resort to hybrid schemes of OMA and NOMA. In this case, the available resources are first split into several identical and orthogonal sub-channels via OMA techniques, e.g., FDMA, TDMA, OFDMA, etc. Thus, users in different sub-channels are orthogonal and can be decoded independently. Moreover,

each sub-channel can be shared with at most two users to reduce detection and decoding complexity.

The concept of Software Defined Multiple Access (SoDeMA) is a flexible and adaptive configuration of several multiple access schemes, orthogonal and non-orthogonal. These schemes will co-exist to satisfy different requirements of diverse services and heterogeneous applications. Real-time video services will need OMA schemes to guarantee their high data rate requirements. On the other hand, mobile social applications will need NOMA for the support of massive connectivity and short packet transmissions.

These schemes will also co-exist to offer trade-offs between performance and complexity. For example, power-domain NOMA and SIC, when feasible, offer low complexity receivers. On the other hand, code-domain NOMA and near-optimal iterative joint decoding can offer high data rates and high reliability. Furthermore, for a large number of users with similar channel gains, code domain NOMA such as LDS-OFDM or MUSA are good choices as they do not require the design of a different codebook per user.

2.4.5 Standardization Activities

Many industry parties such as CATR, CHTTL, HTC, Huawei, HiSilicon, ITRI, MediaTek, NTT DOCOMO, OPPO, and Sony, are involved in investigating the performance of NOMA and the possible standardization in future 3GPP LTE [82] as well as 5G. In fact, 3GPP is conducting a study on NOMA titled Multi User Superposition Transmission (MUST) which is focused on non-orthogonal transmission schemes, advanced processing, detection and decoding at the receiver side, advanced signalling schemes, etc. Achievable cell-average and cell-edge throughput gains are approximated to be as high as 20 percent of existing values. Moreover, 3GPP LTE Release 14 has recently approved a new work item of downlink MUST, whose main goal is to identify the necessary means for enabling downlink intra-cell MUST for the Physical Downlink Shared Channel (PDSCH) in LTE.

Chapter 3

Design of SINR-Based Random Access using Codes-On-Graph

3.1 Chapter Introduction

3.1.1 Chapter Overview

In RA, multiple users transmit to a common AP in different time-frequency resources according to predefined probabilities. When more than one user transmits in the same time-frequency resource, a collision occurs, leading to the loss of these users' information, and users usually need to retransmit their information. There has been a significant amount of work in this field, and various access protocols have been developed to increase access efficiency or to minimize error probabilities. We refer the readers to Section 2.2 for a summary of the main works. However, the simplistic clean packet model[1] discussed therein does not distinguish between different levels of interference but rather treats it as a binary parameter.

Alternatively, when regarding the interference as the accumulation of the power received from other users that are concurrently transmitting, the SINR proves to be a

[1] A clean packet refers to an interference free packet, and the clean packet model is another term for the collision channel model.

more accurate metric for analyzing and designing random access [15]. In this case, a collision is defined as the event of a user's SINR's level dropping below a predefined threshold. By relaxing the clean packet constraint, higher system loads can be achieved for the same power and performance requirements. Previous works have applied the SINR model to a set of ad hoc transmitter/receiver pairs and often refer to it as the physical interference model [83, 84, 85, 86]. However, optimal solutions are shown to be prohibitively complex with no extension to RA.

Motivated by the above, in the first part of this chapter, we investigate the performance of SINR-based RA employing SIC to recover the users' information at the AP. Our objective is to maximize the system load. At each iteration of the SIC process, we assume that a user's information is successfully recovered if its updated SINR is above a predetermined threshold [87]. Following on that assumption, the analogy to the iterative recovery process of codes-on-graph for the BEC no longer holds as having degree-one nodes at each iteration of the SIC process is not a strict requirement for successful recovery in our model; thus, we cannot use the conventional AND-OR tree analysis adopted in [41]. To solve this problem, we develop a new message passing algorithm of the SIC process along with a tree-based analytical framework to evaluate the maximum achievable system load. We show that our work is a generalization of the AND-OR tree-based analysis.

In the second part of this chapter, we investigate the application of this scheme in a CRN setting. Cognitive radio has been recognized as a promising technology to achieve the efficient utilization of the radio spectrum. In CRNs, unlicensed SUs are allowed access to the radio spectrum owned by the licensed PUs, provided that the PUs are guaranteed a certain level of protection. Optimal resource allocation algorithms i.e., channel and power allocation, among the SUs that maximize their data rates or minimize their transmit power requirements have been well-investigated for multiple scenarios and are known to be NP-hard. Accordingly, numerous suboptimal algorithms for resource allocation have been proposed for both downlink and uplink CR transmissions [88, 89]. However, these approaches do not scale well as the number of users in the network increases and their activity becomes more dynamic.

To overcome these problems, RA protocols provide a simple solution that significantly reduces processing and signalling overhead. For example, in [90], authors proposed another random access approach where the Cognitive Base Station (CBS) predetermines a certain transmission probability and makes it known to all the SUs. The PUs' transmissions are fixed whereas the SUs transmissions are randomized according to the assigned transmission probability. It is shown that such a simple random transmission can offer significant improvements in performance, in certain cases, for both the PUs and SUs, compared to fixed transmissions. It is argued that, from a design point of view, controlling the probabilities is easier than controlling the power. However, the authors only considered a very single case of a single channel and no analysis was done to derive the design criteria for choosing the optimal transmission probability.

3.1.2 Chapter Contributions

The main contributions of this chapter are summarized below.

Tree-based Analysis of SINR-based Random Access

We revisit random access for wireless systems with SIC employed at the AP. We consider an asymptotically large number of devices that transmit over a large number of orthogonal sub-channels. In each transmission block, each device chooses a degree d, where d is a random variable that follows a predefined degree distribution $\Omega(x)$. Then, devices transmit in d sub-channels chosen uniformly at random. Specifically, we consider SINR-based RA where it is assumed that a device's information can be recovered successfully at a given iteration of the SIC process when its updated SINR is above a predetermined threshold. We develop a generalized analytical framework based on the codes-on-graph representation to track the evolution of error probabilities in each iteration of the SIC process. We compare our approach to the conventional RA employing SIC which assumes that only clean, interference-free transmissions can be recovered successfully. This clean packet model relies on having

time slots with a single device's transmission at each iteration of the SIC process. It was shown to be analogous to the iterative recovery process of codes-on-graph for the BEC, thus, allowing the direct application of the AND-OR tree analysis. We show that the clean packet model is a special case of our more generalized tree-based analytical framework.

Degree Distribution Optimization of SINR-based Random Access

Based on the derived analytical framework, we optimize the design parameters evolution strategies, e.g., differential evolution, such that the error probabilities are minimized. Our numerical results show that the SINR metric allows for the support of a larger number of devices, under the same power constraints and reliability requirements.

Design of SINR-based Random Access in Cognitive Radio Networks

We investigate the application of SINR-based RA in CRNs. The CBS allows the SUs to reuse the sub-channels of the PUs provided that the interference of the SUs to the PUs is below a predetermined threshold. PUs transmit over a fixed set of channels with fixed transmission powers that are scheduled by the CBS. Once the signals of the SUs and PUs are received, CBS then implements SIC to recover both the SUs' and PUs' signals. In the signal recovery, we assume that the PUs' signals can be recovered if the Interference Power (IP) of the SUs to the PUs is below a predetermined threshold. On the other hand, we assume that the SUs' signals can be recovered if its received SINR is above a predetermined threshold. We formulate a new optimization problem to find the optimal degree distribution function that maximizes the probability of successfully recovering the signals of an SU in the SIC process under the SINR constraints of the SUs while satisfying the IP constraints of the PUs. Simulation results show that our proposed design can achieve higher success probabilities and a lower number of transmissions in comparison with conventional

Table 3.1 – Notation Summary

Notation	**Description**
K	Number of active MTC devices
$\mathcal{K}$	Set of active MTC devices
D_k	k^{th} MTC devices
N	Number of orthogonal sub-channels
CH_n	n^{th} orthogonal sub-channel
$\mathcal{N}_k$	Set of channels chosen by D_k
h_{kn}	Channel gain between D_k and the AP over CH_n
P_{nk}	Transmit power of D_k over the CH_n
$\gamma^{(k,\ell)}$	SINR of D_k after ℓ iterations of SIC
$\gamma_{\mathrm{th}}^{(k)}$	Required SINR threshold for D_k
P_o	Received power level per device per sub-channel
$\Omega(x)$	Generator polynomial of the degree distribution of the sub-channels
Ω_i	Probability of a sub-channel with degree equal to i
$\bar{\Omega}$	Average degree of a sub-channel
$\Lambda(x)$	Generator polynomial of the degree distribution of the devices
Λ_i	Probability of a device with degree equal to i
$\bar{\Lambda}$	Average degree of a device

schemes, thus, significantly improving the signal recovery performance and reducing energy consumption.

3.1.3 Chapter Outline

The rest of this chapter is organized as follows. Section 3.2 presents the system model. In Section 3.3, we describe the transmission scheme and the SIC process. In Section 3.4, we analyze the system performance in an asymptotic setting and formulate our degree distribution optimization problem. Section 3.5 is dedicated to the application of SINR-Based RA to CRN. It covers the system modelling, analysis, optimization and comparison to conventional RA. Numerical results are shown in Section 3.6. Finally, Section 3.7 concludes the chapter.

The notations used in this chapter are summarized in Table 3.1 for quick reference.

3.2 System Model

We consider a multiple access channel with a single AP and a set of active devices $\mathcal{K}$ of size K. Devices are denoted by $\{\mathrm{D}_1,...,\mathrm{D}_K\}$. We consider a total number of N orthogonal sub-channels denoted by $\{\mathrm{CH}_1,...,\mathrm{CH}_N\}$, over which the devices are allowed to transmit their information.

Channels are assumed to be reciprocal and block fading; that is, we assume that the channel coefficients remain constant for the whole block length but vary independently from one block to the other. Let $\mathbf{y}=[y_n]_{1\leq n\leq N}$ denote the received signal vector at the AP and is given by:

$$\mathbf{y} = \mathbf{H}\mathbf{x} + \mathbf{e}, \tag{3.1}$$

where $\mathbf{H} = [h_{k,n}]_{1\leq k\leq K, 1\leq n\leq N}$ is the channel matrix, where $h_{k,n}$ denotes the channel gain between D_k and the AP over the n^{th} sub-channel. $\mathbf{x} = [x_k]_{1\leq k\leq K}$ is the transmitted signal vector, and $\mathbf{e}$ is the AWGN vector with zero mean and variance $\sigma^2 I_n$. Let $\mathcal{N}_k$ denote the set of sub-channels that have been selected by U_k. We denote by $P_{n,k}$ the transmit power of D_k over the n^{th} sub-channel. Then, $P_k = \sum_{n=1}^{N} P_{k,n}$ is the total transmit power of D_k, where $P_{k,n} = 0$ for $n \notin \mathcal{N}_k$.

For simplicity, we also assume that each device transmits the same message with equal power over all chosen sub-channels. This approach has been justified in [91] for uplink networks. The system can, then, be represented by a bipartite graph shown in Figure 3.1, where the devices and sub-channels are shown by circles and squares, representing the variable nodes and check nodes, respectively.

3.3 SINR-Based Random Access Scheme

In this section, we describe the transmission scheme for the previously described system. First, we introduce the main design parameter, namely the degree distribution $\Omega(x)$. Then, we provide a novel message passing representation of the SIC process in SINR-based RA.

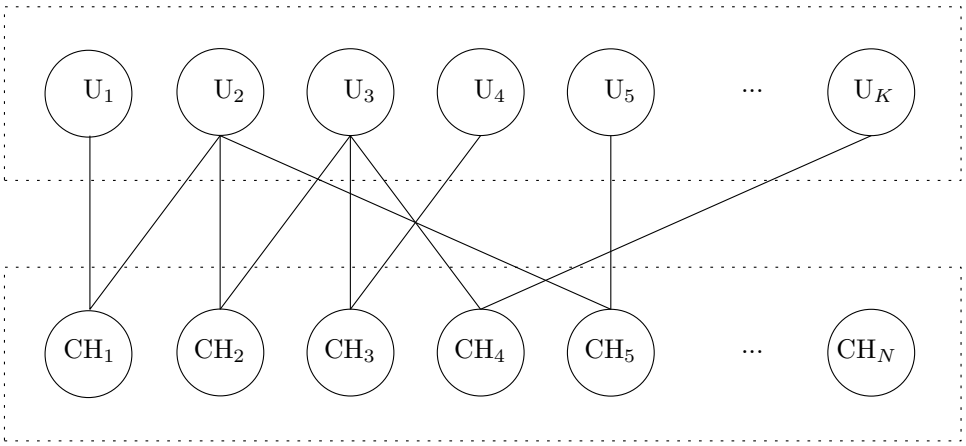

Figure 3.1 – Bipartite Graph Representation of SINR-Based Random Access.

3.3.1 Degree Distribution

For a given transmission block, each device chooses a degree d obtained from a pre-defined Probability Density Function (PDF) whose generator polynomial is defined as $\Omega(x) := \sum_i \Omega_i x^i$, where Ω_i is the probability of the degree being i. We refer to $\Omega(x)$ as the degree distribution. Then, the device chooses d sub-channels uniformly at random to transmit over. The degree of each device is defined as the number of edges connected to that device in the bipartite graph (Figure 3.1).

Since the sub-channels are selected by devices uniformly at random, the degree of each sub-channel, defined as the number of devices transmitting in that sub-channel, follows the Poisson distribution [44]. For a degree d sub-channel, Λ_i denotes the probability of $d = i$. Then, Λ_i is given by:

$$\Lambda_i = e^{-\alpha}\frac{\alpha^i}{i!}, \tag{3.2}$$

which can be also expressed in generator polynomial form as $\Lambda(x) = \exp^{\alpha(x-1)}$, in the asymptotic setting [44], where $\alpha = \frac{K}{N}\bar{\Omega}$ and $\bar{\Omega} = \sum_d d\Omega_d$ is the average device degree.

3.3.2 Successive-Interference Cancellation

We assume that the AP employs SIC to recover the transmitted information of each device. For a given iteration of the SIC process, a device's information can be successfully recovered provided that its received SINR is above a predetermined threshold $\gamma_{\text{th}}^{(k)}$, where $k \in \mathcal{K}$ [87]. Once a device is recovered at the AP, the interference caused by that device can be completely removed from its selected sub-channels, which in turn increases the SINRs of the remaining devices in the following iterations.

Let $\mathcal{K}_n$ denote the set of devices transmitting in the n^{th} sub-channel. Let $\mathcal{U}^{(\ell)}$ denote the set of all successfully recovered devices after ℓ iterations of the SIC process at the AP. The total SINR of U_k after ℓ iterations of the SIC process is denoted by $\gamma^{(k,\ell)}$ and can be calculated as follows:

$$\gamma^{(k,\ell)} = \sum_{n \in \mathcal{N}_k} \frac{|h_{k,n}|^2 \frac{P_k}{|\mathcal{N}_k|}}{\sum\limits_{i \in \mathcal{K}_n - \mathcal{U}_\ell, i \neq k} |h_{i,n}|^2 \frac{P_i}{|\mathcal{N}_i|} + \sigma^2}, \quad \text{for} \quad k \notin \mathcal{U}^\ell, \tag{3.3}$$

where $|\mathcal{X}|$ is cardinality of the set $\mathcal{X}$. The AP can recover U_k's information in the ℓ^{th} iteration of the SIC process, if and only if $\gamma^{(k,\ell)} \geq \gamma_{\text{th}}^{(k)}$ for $k \in \mathcal{K}$, which will happen for a certain level of interference.

For a large number of devices, centralized power control is not feasible. Instead, devices are assumed to tune their transmit powers adaptively, as required, while having the received power level known at the AP [92]. For ease of analysis, we assume that the received power level of each device over each of its chosen sub-channels is P_o, with the extension to the general case being straightforward. Moreover, we assume that all devices share the same SINR threshold γ_{th}. Thus, the interference caused to U_k's transmission over the n^{th} sub-channel can be expressed as $d_n^{(k,\ell)} P_o$, where $d_n^{(k,\ell)}$ is a random variable that represents the number of unrecovered devices in the ℓ^{th} iteration that are transmitting in the n^{th} sub-channel, other than k. Thus, from Equation 3.3, $d_n^{(k,\ell)} = |\mathcal{K}_n - \mathcal{U}^{(\ell)}| - 1$, for $n \in \mathcal{N}_k$.

For a given device U_k, we define $\mathbf{d}^{(k,\ell)} = [d_n^{(k,\ell)}]$, for $n \in \mathcal{N}_k$, as its observation vector. Moreover, we define a search set $\mathbf{V}^{(i)}$ as the set of all vectors $\mathbf{v}$ that can satisfy the

SINR threshold for a degree i device. It can, then, be found as follows:

$$\mathbf{V}^{(i)} = \{(v_1, v_2, ..., v_i) | \sum_{n=1}^{i} \frac{P_o}{v_n P_o + \sigma^2} \geq \gamma_{\text{th}}\}. \tag{3.4}$$

Thus, the AP can recover U_k's information if and only if the observation vector $\mathbf{d}^{(k,\ell)}$ belongs to $\mathbf{V}^{(|\mathcal{N}_k|)}$. Thus, the probability of recovering U_k's information with error at the ℓ^{th} iteration of the SIC process can be expressed as follows:

$$q^{(k,\ell)} = 1 - \Pr(\mathbf{d}^{(k,\ell)} \in \mathbf{V}^{(|\mathcal{N}_k|)}). \tag{3.5}$$

3.4 Iterative Convergence Analysis

In this section, we first propose a novel message representation of the SIC algorithm in SINR-based RA. Then, we propose an iterative tree-based analytical framework for analyzing the error probability of the proposed random transmission scheme. We show that this approach is more general compared to the existing AND-OR tree analysis [45], which has been widely used for analyzing iterative decoding of codes-on-graph for the BEC [43, 44]. Finally, we show that the SINR model allows for lower error probabilities in comparison to the clean packet model for the same system load.

3.4.1 Representing SIC as a Message Passing Algorithm

Let us first rephrase the SIC process for the proposed approach as a message passing algorithm. For a given system with K devices and N sub-channels, we first construct the corresponding bipartite graph. Let $\mathcal{N}_i$ denote the set of variable nodes connected to check node i and $\mathcal{N}_i \backslash j$ denote the set of variable nodes connected to check node i except variable node j. Let $\mathcal{M}_j$ denote the set of check nodes connected to variable node j. The message passed from check node i to variable node j in the ℓ^{th} iteration of the SIC process is denoted by $m_{i,j}^{(\ell)}$, and the message passed from variable node j to check node i in the ℓ^{th} iteration of the SIC process is denoted by $n_{j,i}^{(\ell)}$. In what

follows, we provide a novel message passing representation of the SIC process.

- *Initialisation:* Each variable node is assigned with a value of 0, i.e., $n_{i,j}^{(0)} = 0$ for $1 \leq i \leq K$ and $1 \leq j \leq N$.
- For each iteration $\ell \geq 1$:

1. *Check-to-Variable node update:* The message passed from check node i to variable node j in the ℓ^{th} iteration of the SIC process, denoted by $m_{i,j}^{(\ell)}$, is given by:

$$m_{i,j}^{(\ell)} = \sum_{j' \in \mathcal{N}_i \setminus j} (1 - n_{j',i}^{(\ell-1)}). \tag{3.6}$$

2. *Variable-to-Check node update:* The message passed from variable node j to check node i in the ℓ^{th} iteration of the SIC process, denoted by $n_{j,i}^{(\ell)}$, is given by:

$$n_{j,i}^{(\ell)} = \begin{cases} 1 & \text{if } (m_{i_1,j}^{(\ell)}, ..., m_{i_{|\mathcal{M}_j|},j}^{(\ell)}) \in \mathbf{V}^{(|\mathcal{M}_j|)}, \\ 0 & \text{Otherwise,} \end{cases}$$

   where $\mathbf{V}^{(|\mathcal{M}_j|)} = \{\mathbf{v} | \sum_{i=1}^{|\mathcal{M}_j|} \frac{P_o}{v_i P_o + \sigma^2} \geq \gamma_{\text{th}}\}$.

3. Repeat steps 1 and 2 until all variable nodes have received message 1 or a predetermined number of iterations has passed.

In the above representation, it is important to note that each variable node sends a message to its neighboring check node if and only if its total SINR is larger than the SINR threshold. Moreover, a check node sends a message v to its neighboring variable node if and only if that check node is connected to v unrecovered variable nodes from the remaining variable nodes. This is exactly what happens in the SIC process, as the message of value 1 from variable to check node means that the variable node has been recovered and its corresponding edges have been removed from the bipartite graph.

3.4.2 Generalized Tree Analysis

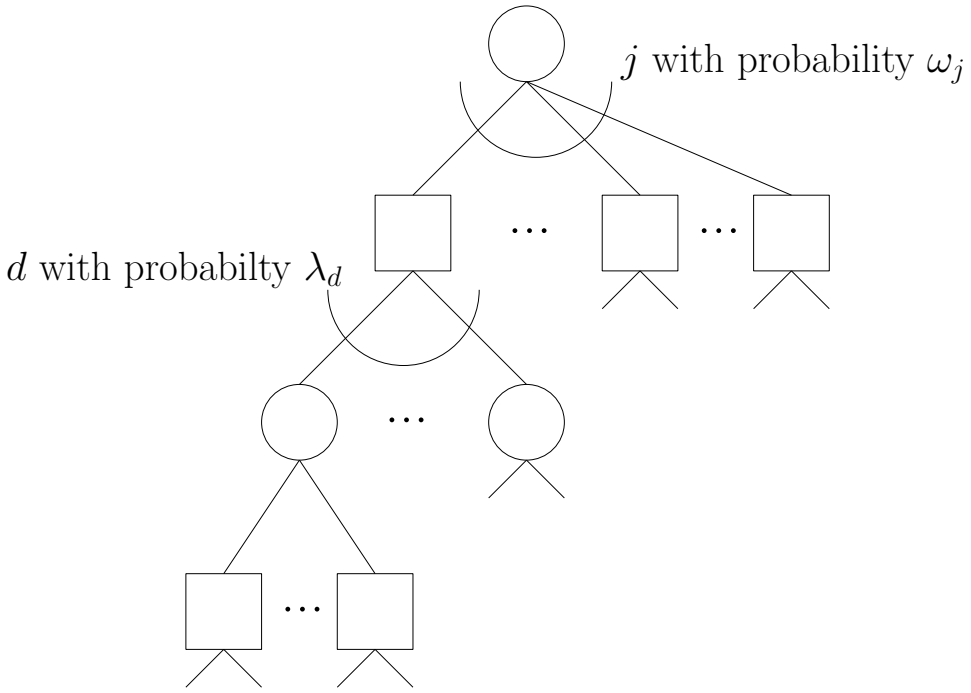

Figure 3.2 – C-V tree $\mathcal{T}_\ell$.

Here, we propose a general framework to analyze SINR-based RA employing SIC. As in the conventional iterative analysis techniques (AND-OR tree technique), we first represent the graph as a tree. Let $\mathcal{T}$ denote the bipartite graph corresponding to the multi-user system at the AP. Consider a random subgraph $\mathcal{T}_\ell$ as follows: Choose an edge (v, w) uniformly at random. Subgraph $\mathcal{T}_\ell$ is the graph induced by variable node v and all the neighbors of v within distance 2ℓ after removing the edge (v, w). As shown in [44], $\mathcal{T}_\ell$ is a tree asymptotically, so we refer to $\mathcal{T}_\ell$ as a C-V tree, where C and V nodes refer to check and variable nodes, respectively.

Figure 3.2 illustrates the C-V tree of depth 2ℓ. Nodes at depth i have children at depth $i + 1$. The root of the tree is a variable node at depth 0. More specifically, variable nodes are located at depths $0, 2, ..., 2\ell$ and have j children with probability ω_j; on the other hand, check nodes are located at depths $1, 3, ..., 2\ell - 1$ and have j children with probability λ_j.

From a graphical perspective, ω_j is the probability that a uniformly chosen edge is connected to a variable node of degree j. Similarly, λ_j is the probability that a uniformly chosen edge is connected to a degree-j check node. It follows that the degree distributions from an edge perspective can be defined as follows:

$$\omega(x) \triangleq \frac{1}{\bar{\Omega}} \sum_i i\Omega_i x^{i-1}, \quad \lambda(x) \triangleq \frac{1}{\alpha} \sum_i i\Lambda_i x^{i-1}. \tag{3.7}$$

Since a subgraph expanded from each variable node is asymptotically a tree, the message passing between variable and check nodes in the bipartite graph can be seen as the message passing between V and C nodes in the tree. In each iteration of the SIC process, V nodes at depth $2i$ pass messages to their C node parents at depth $2i-1$, which, in return, calculate messages of their own to be passed to their V node parents at depth $2i-2$. Starting from depth 2ℓ, we see that ℓ iterations are required to have the messages reach the root of the tree. Then, the error probability is defined as the probability that the messages received at the root do not satisfy the SINR threshold.

The following lemma gives the recovery error probability of the SIC process in SINR-based RA.

Lemma 1. *For SINR-based RA, let q_ℓ denote the probability that a device cannot be recovered at the ℓ^{th} iteration of the SIC process with the SINR threshold γ_{th}. Then, q_ℓ is expressed as follows:*

$$q_\ell = 1 - \sum_{i=0}^{N-1} \omega_i \sum_{\boldsymbol{v} \in \boldsymbol{V}^{(i+1)}} \prod_{j=1}^{i+1} \sum_{d=v_j}^{K-1} \lambda_d \binom{d}{v_j} (1-q_{\ell-1})^{d-v_j} q_{\ell-1}^{v_j}, \tag{3.8}$$

where $q_0 = 1$ and $\boldsymbol{V}^{(i+1)} = \{\boldsymbol{v} | \sum_{j=1}^{i+1} \frac{P_o}{v_j P_o + \sigma^2} \geq \gamma_{th}\}$.

The proof of this lemma is provided in Section A.1. As outlined, this work is a generalisation of the conventional AND-OR tree analysis. This is shown in the following proposition.

Proposition 3.1. *Conventional RA employing SIC, where only clean packets can be successfully recovered, is a special case of our approach with the search set* $\boldsymbol{V}_{conv}$ *expressed below.*

$$\boldsymbol{V}_{conv}^{(i)} = \{\boldsymbol{v} | \exists j, v_j = 0, 1 \leq j \leq i\},$$

and the following recovery error probability q_ℓ *[41]:*

$$q_\ell = \omega(1 - \lambda(1 - q_{(\ell-1)})). \tag{3.9}$$

The proof of this proposition is provided in Section A.2.

Proposition 3.2. *The recovery error probability of the proposed approach with the SINR threshold* $\gamma_{th} \leq \frac{P_o}{\sigma^2}$ *can be rewritten as follows:*

$$q_\ell = \omega(1 - \lambda(1 - q_{(\ell-1)})) - \sum_{i=0}^{N-1} \omega_i \sum_{\boldsymbol{v} \in \boldsymbol{V}_0^{(i+1)}} \prod_{j=1}^{i+1} \sum_{d=v_j}^{K-1} \lambda_d \binom{d}{v_j} (1 - q_{\ell-1})^{d-v_j} q_{\ell-1}^{v_j},$$

where $\boldsymbol{V}_0^{(i+1)} = \{\boldsymbol{v} | \sum_{j=1}^{i+1} \frac{P_o}{v_j P_o + \sigma^2} \geq \gamma_{th}, v_i > 0\}$.

This proposition can be easily proved by using the results of Proposition 3.1 to simplify the results of Lemma 1. This proposition shows that the SINR model achieves a lower recovery error probability compared to conventional SIC techniques based on the clean packet model.

3.4.3 Degree Distribution Optimization

The system load C is defined as the ratio $\frac{K}{N}$, and the maximum system load C^* is restricted by the following condition for convergence:

$$C^* = \max_C \{q_i < q_{i-1}\}, \text{ for } i > 0. \tag{3.10}$$

Table 3.2 – Optimal Degree Distributions for SINR-based RA

P_o	1 dB	3 dB	5 dB
γ_{th}	1 dB	3 dB	5 dB
Ω_1	0.0000	0.0000	0.0000
Ω_2	0.3047	0.5631	0.5102
Ω_3	0.1491	0.0436	0.0000
Ω_4	0.5462	0.3933	0.4898
C^*	1.3740	1.0450	0.8680
Ω	3.2415	3.2235	2.9796

Given the received power levels and the SINR thresholds, we formulate an optimization problem to find the degree distribution that can maximize the achievable system load. It can be written as:

$$\begin{aligned} &\max_{\Omega(x)} && C^*, \\ &\text{s.t.} && \text{(i) } 0 \leq \Omega_i \leq 1, && \forall 1 \leq i \leq d_m. \\ & && \text{(ii) } \sum_{i=1}^{d_m} \Omega_i = 1. \end{aligned}$$

Using differential evolution [93], we can find the optimal degree distribution that maximizes C^* for any P_o and γ_{th}. In brief, differential evolution is a numerical optimization method that solves the formulated problem in an iterative fashion. In each iteration, a number of candidate solutions are produced and are given weights representing a measure of quality. The algorithm aims at improving the measure of quality of the generated candidates in each iteration. For a large number of iterations, the algorithm will converge to either an optimal or sub-optimal solution. However, although it cannot guarantee optimality, we have found it to be a valuable tool for our problem.

We denote by d_m the maximum allowed degree such that $\Omega_i = 0$ for $i > d_m$. Table 3.2 shows the optimal degree distributions for $d_m = 4$ and different SINR thresholds. For fair comparison between the SINR model and the clean packet model, we only considered the case where $P_o = \gamma_{\text{th}}$. The evolution of the error probabilities in each iteration of the SIC process according to Lemma 1 is shown in Figure 3.3.

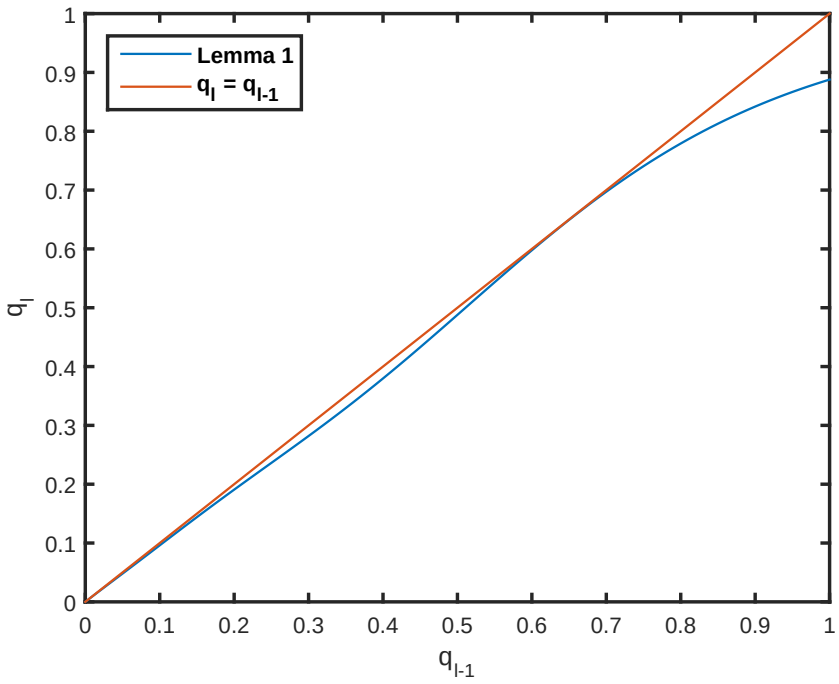

Figure 3.3 – Evolution of the probability q according to the equation derived in Lemma 1 for $P_o = \gamma_{\text{th}} = 3$ dB, $\Omega(x) = 0.5631x^2 + 0.0436x^3 + 0.3933x^4$, and $C = 1.045$.

From Proposition 3.2, we expect the performance of the SINR model to be equivalent or better than the clean packet model depending on the values of P_o and γ_{th}. Interestingly, we find that for $P_o = \gamma_{\text{th}} = 5$ dB, the optimal degree distribution and C^* are the same as that found in [41] for $d_m = 4$. That is because $V_0^{(i)}$ yields an empty set $\forall\ i$, and the achievable performance is expected to be the same for both models in that case.

3.5 Application of SINR-based Random Access to Cognitive Radio Networks

In the second part of this chapter, we investigate the application of SINR-based RA to CRNs. For clarity purposes, we have provided another table of notations for this

Table 3.3 – Section 3.5 Notation Summary

Notation	Description
K_p	Number of active PUs
$\mathcal{K}_p$	Set of active PUs
K_s	Number of active SUs
PU_k	k^{th} PU
SU_k	k^{th} SU
$\mathcal{K}_s$	Set of active SUs
$\mathcal{N}_p^{(k)}$	Set of sub-channels allocated to PU_k
$\mathcal{N}_s^{(k)}$	Set of sub-channels chosen by SU_k
N	Number of orthogonal sub-channels
CH_n	n^{th} orthogonal sub-channel
h_{kn}	Channel gain between D_k and the AP over CH_n
$P_{k,n}$	Transmit power of PU_k over the CH_n
$Q_{k,n}$	Transmit power of SU_k over the CH_n
$\gamma_p^{(k,\ell)}$	SINR of PU_k after ℓ iterations of SIC
$\gamma_s^{(k,\ell)}$	SINR of SU_k after ℓ iterations of SIC

section, namely, Table 3.3.

3.5.1 CRN System Model

We consider an uplink CRN, including a CBS, a set of K_p active PUs, denoted by $\mathcal{K}_p$, and a set of K_s active SUs denoted by $\mathcal{K}_s$. There are in total N orthogonal sub-channels of equal bandwidth in the network. Channels are assumed to be reciprocal and block fading; that is, we assume that the channel coefficients remain constant for the whole transmission block but vary independently from one block to the other. Let y_n denote the received signal vector at the CBS over the n^{th} sub-channel, where $1 \leq n \leq N$. Then, it can be expressed as follows:

$$y_n = \sum_{k \in \mathcal{K}_p} g_{k,n} x_{k,n} + \sum_{i \in \mathcal{K}_s} h_{i,n} u_{i,n} + e_n, \tag{3.11}$$

where $g_{k,n}$ is the channel gain between PU_k and the CBS over the n^{th} sub-channel, and $h_{k,n}$ is the channel gain between SU_k and the CBS over the n^{th} sub-channel. $x_{k,n}$ and $u_{i,n}$ are the transmitted signals of each of the PUs and SUs to the CBS, over the

n^{th} sub-channel. e_n is a AWGN random variable with zero mean and variance σ_e^2.

Each PU is allocated one distinct set of sub-channels. We denote by $\mathcal{N}_p^{(k)}$ the set of $N_p^{(k)}$ sub-channels allocated to PU_k. We denote by $\mathcal{N}_s^{(k)}$ the set of $N_s^{(k)}$ sub-channels chosen by SU_k. Then, $x_{k,n} = 0$ for $n \notin \mathcal{N}_p^{(k)}$, and $u_{k,n} = 0$ for $n \notin \mathcal{N}_s^{(k)}$. Moreover, we denote by $Q_{k,n}$ and $P_{k,n}$ the power of $x_{k,n}$ and $u_{k,n}$, respectively.

3.5.2 CRN Transmission Scheme

For a given transmission block, each SU chooses a random degree d obtained from a predefined degree distribution whose generator polynomial is given as $\Omega(x) = \sum_i \Omega_i x^i$, where Ω_i is the probability that $d = i$. Then, the SU chooses d sub-channels uniformly at random to transmit over. We define a $K_s \times N$ random channel access matrix $\mathbf{A}$ with integer elements $a_{k,n} \in \{0, 1\}$, where $a_{k,n} = 1$ means SU_k is transmitting in sub-channel n, and $a_{k,n} = 0$ means SU_k is not transmitting in sub-channel n. Thus, it is easy to show that the elements of $\mathbf{A}$ are i.i.d. Bernoulli random variables with a success probability of $\frac{1}{N}\bar{\Omega}$, where $\bar{\Omega}$ is the average degree and is given by $\sum_i i\Omega_i$. Then, we can represent the probabilistic random transmission scheme by a bipartite

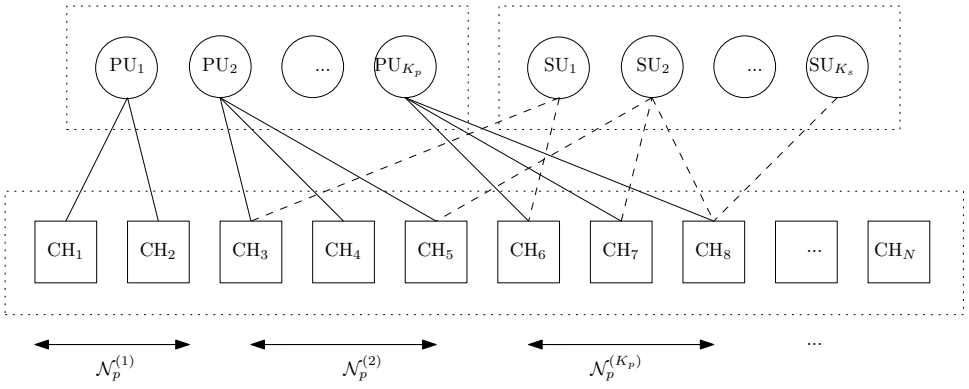

Figure 3.4 – Bipartite Graph Illustration of the SINR-Based RA Scheme in a CRN

graph as shown in Figure 3.4. The PUs and SUs are shown by circles and referred

to as variable nodes while the sub-channels $[\mathrm{CH}_i]_1 \leq i \leq N$ are shown by squares and referred to as check nodes. The number of edges connected to each variable node corresponds to the number of sub-channels it is allocated, and it is called the degree of the respective variable node. The solid edges represent the transmissions of the PUs whose number is assumed to be fixed, e.g., PU_1 is of degree 2 in Figure 3.4. On the other hand, the dashed edges represent the transmissions of the SUs whose number is a random variable with a distribution pre-determined by the CBS.

The CBS employs SIC to recover each device's signal. Each device is assumed to transmit the same signals over its respective sub-channels. The CBS can, then, combine the received transmissions of each device over all respective sub-channels using Maximal Ratio Combining (MRC), and the overall received SINR at the CBS can be represented as the sum of all individual SINRs. Note that the CBS is assumed to have the perfect knowledge of the PUs' channel state, transmit power and allocated sub-channels. We also assume that the CBS first attempts to recover the signals of the PUs. The maximum achievable rate of PU_i, where $1 \leq i \leq K_p$, is shown below:

$$R_p^{(i)} = \frac{N_p^{(i)}}{N} \log \left(1 + \sum_{n \in \mathcal{N}_p^{(i)}} \gamma_{p,n}^{(i)} \right), \tag{3.12}$$

where

$$\gamma_{p,n}^{(i)} = \frac{|g_{i,n}|^2 Q_{i,n}}{\sum_{k=1}^{K_s} a_{k,n} |h_{k,n}|^2 P_{k,n} + \sigma_e^2}. \tag{3.13}$$

We denote by $I_{p,n}^{(i)} = \sum_{k \in \mathcal{K}_s} a_{k,n} |h_{k,n}|^2 P_{k,n}$ the interference power caused to PU_i's transmission over the n^{th} sub-channel, where $n \in \mathcal{N}_p^{(i)}$. Thus, the total interference caused by the SUs to PU_i can be expressed as $I_p^{(i)} = \sum_{n \in \mathcal{N}_p^{(i)}} I_{p,n}^{(i)}$. The signals of PU_i can be successfully recovered provided that $I_p^{(i)}$ is below the threshold $I_{\mathrm{th}}^{(i)}$.

Without loss of generality, we assume the SUs' signals are recovered through the SIC process according to their received SINR, in an ascending order. More specifically, we assume that the SINR of SU_k is larger than that of SU_{k-1}, for $1 \leq k \leq K_s$. Assuming the signals of the first $i-1$ SUs have been successfully recovered, the maximum

achievable rate of SU_i, where $1 \leq i \leq K_s$, is shown below:

$$R_s^{(i)} = \frac{N_s^{(i)}}{N} \log \left(1 + \sum_{n \in \mathcal{N}_s^{(i)}} \gamma_{s,n}^{(i)} \right), \tag{3.14}$$

where

$$\gamma_{s,n}^{(i)} = \frac{a_{i,n}|h_{i,n}|^2 P_{i,n}}{\sum_{k=i+1}^{K_s} a_{k,n}|h_{k,n}|^2 P_{k,n} + \sigma_e^2}. \tag{3.15}$$

Thus, we can express the total SINR of SU_i as $\gamma_s^{(i)} = \sum_{n \in \mathcal{N}_s^{(i)}} \gamma_{s,n}^{(i)}$. The signals of SU_i can be successfully recovered provided that their received SINR $\gamma_s^{(i)}$ at the i^{th} iteration of SIC is above the threshold $\gamma_{\mathrm{th}}^{(i)}$.

It will be shown later that the design problem is dependent on the SUs' received power rather than transmit power. Assuming that the SUs are able to estimate their channel gains from the downlink given the reciprocity of the channel, the CBS needs to broadcast the received power constraints only, imposed on each sub-channel on a per device basis. The SUs can, then, adaptively tune their power as necessary. Accordingly, we define a power vector $\mathbf{p} = [P_n]_{1 \leq n \leq N}$, where P_n is the received power constraint imposed on the n^{th} sub-channel on a per device basis.

3.5.3 Asymptotic Performance Analysis of SINR-Based RA in CRNs

In this section, we analyze the relationship between the system constraints (I_{th} and γ_{th}) and the different system metrics (N, K_p and K_s). We formulate an optimization problem to find the degree distribution that can maximize this probability of successfully recovering the signals of the SUs for a given setup.

Probability Density Function of the IP

Let us first calculate the power of interference introduced to the PUs.

Lemma 2. *In an asymptotically large network ($N \to \infty$, $K_s \to \infty$), the probability density function of the total interference power induced over the sub-channels of PU_k, $\forall k \in \mathcal{K}_p$, follows the Poisson distribution below:*

$$Pr(I_p^{(k)} = iP_o^{(k)}) = e^{-\alpha N_p^{(k)}} \frac{\left(\alpha N_p^{(k)}\right)^i}{i!}, \tag{3.16}$$

where $\alpha = \frac{K_s}{N}\bar{\Omega}$, and $P_n = P_o^{(k)}$ $\forall n \in \mathcal{N}_p^{(k)}$. Its average and standard deviation are given below:

$$\mathbb{E}[I_p^{(k)}] = \alpha N_p^{(k)} P_o^{(k)}, \quad \sigma_{I_p^{(k)}} = \alpha N_p^{(k)} P_o^{(k)}. \tag{3.17}$$

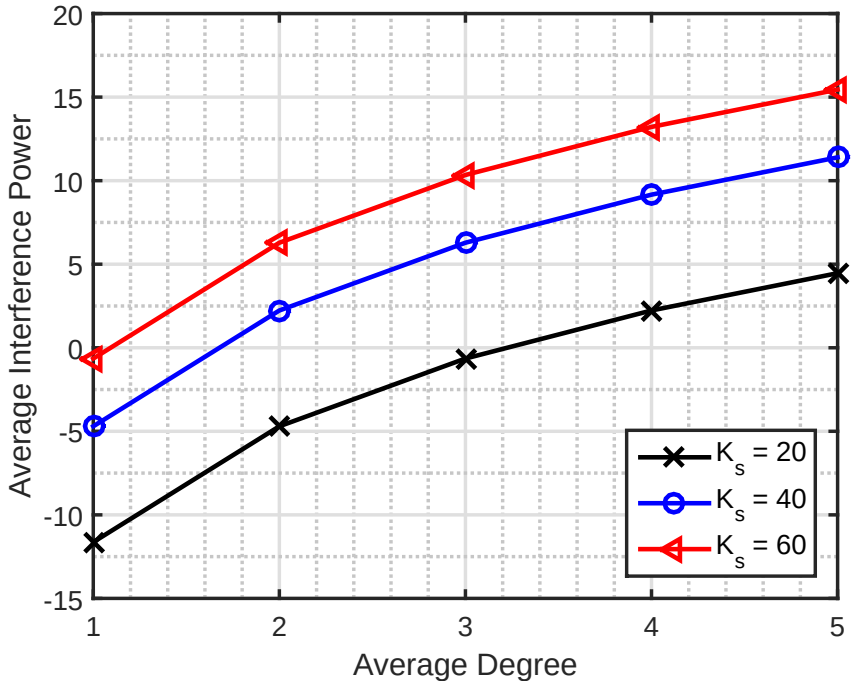

Figure 3.5 – The average interference power for a total of $N = 128$ sub-channels assigned equally to $K_p = 60$ PUs and shared by K_s SUs.

The proof of this lemma is provided in Section A.3. In Figure 3.5, the average IP is shown as a function of the average degree $\bar{\Omega}$ and the number of devices K_s. $P_o^{(k)}$ is set to 0 dB $\forall k \in \mathcal{K}_p$. The average IP per PU is shown to increase with the number of K_s, as expected from Lemma 1. It is worthy of noting that N is fixed for all three

simulations and that the increase in the average IP in fact corresponds to the increase in the ratio $\frac{K_s}{N}$ rather than K_s itself.

Probability of Success of the SUs

As before, the SUs' signals are recovered in an ascending order, based on their received SINR, with $\gamma_s^{(i)} \leq \gamma_s^{(i-1)}$ for $1 \leq i \leq K_s$. Given that the signals of the first $i-1$ SUs have been successfully recovered, we can rewrite Equation 3.15 and express the total SINR of SU_i as follows:

$$\gamma_s^{(i)} = \sum_{n \in \mathcal{N}_s^{(i)}} \frac{P_n}{d_n^{(i)} P_n + \sigma_e^2}, \tag{3.18}$$

where $d_n^{(i)}$ is a random variable that represents the number of devices, other than SU_i, transmitting in the n^{th} sub-channel and whose signals have not been recovered yet. We define $\mathbf{d}^{(i)} = [d_n^{(i)}]_{1 \leq n \leq N_s^{(i)}}$ and refer to it as the observation vector. The CBS can then recover the signals of SU_i, if and only if, $\gamma_s^{(i)} \geq \gamma_{\text{th}}^{(i)}$, which will happen for certain values of $\mathbf{d}^{(i)}$. Let $\mathbf{V}^{(k)}$ denote the set of all vectors $\mathbf{v}$ that can satisfy the SINR constraint for SU_k. It can then be found that:

$$\mathbf{V}^{(k)} = \{(v_1, v_2, ..., v_{N_s^{(k)}}) | \sum_{n \in \mathcal{N}_s^{(k)}} \frac{P_n}{v_n P_n + \sigma_e^2} \geq \gamma_{\text{th}}^{(k)}\}. \tag{3.19}$$

In other words, the CBS can recover the signals of SU_k if and only if the observation vector $\mathbf{d}^{(k)}$ belongs to $\mathbf{V}^{(k)}$. We then have the following proposition:

Proposition 3.3. *For the recovery of the SUs' signals, we assume that the PUs' signals have been successfully recovered and that the SUs' signals are ordered and recovered in an ascending order, based on their received SINR. Let S_i be the event of having $\gamma_s^{(i)} \geq \gamma_{th}^{(i)}$. Then, the probability of successfully recovering the signals of SU_i, through the SIC process, can be calculated as follows:*

$$\begin{aligned} Pr(S_i) &= Pr(\gamma_s^{(i)} \geq \gamma_{th}^{(i)}) \\ &= Pr(\gamma_s^{(i)} \geq \gamma_{th}^{(i)} | S_{i-1}) Pr(S_k) \\ &= Pr(\boldsymbol{d}^{(i)} \in \boldsymbol{V}^{(i)} | S_{i-1}) Pr(S_k), \end{aligned}$$

for $1 \leq i \leq K_s$.

Clean Packet Model

As mentioned before, authors in [41, 94, 95] have implemented the iterative recovery process of codes-on-graph for the BEC in RA schemes. As in Figure 3.4, the system is mapped onto a bipartite graph and the signal recovery is visualized as a message passing algorithm [45]. However, at the receiver side, successful signal recovery can only take place if an interference-free clean packet has been received at the destination.

The observation vector of the 'clean packet' model has the following form for successful signal recovery:

$$\{\mathbf{d}^{(i)} | \exists j, d_j^{(i)} = 0, 1 \leq j \leq i\}.$$

Let us consider the case where the received power of an SU's signal is less than or equal to its SINR threshold. Then, if the received signal is interference-free, it can be successfully recovered in our design. For such a case, the observation vectors of both designs are the same for $d_m = 1$.

On the other hand, from Equation 3.19, we can see that the set of observation vectors that ensure successful recovery will generally be larger for our design; thus, it is expected to provide a higher probability of success. The 'clean packet' model can be seen as a special case of our design. Interestingly, when the SINR threshold is higher than that of the received power per signal for an SU, the 'clean packet' model fails to service any SUs at all. However, for sufficiently high degrees, our approach can still service a significant fraction of the SUs.

Optimization of the Degree Distribution

Given a CRN system of K_p PUs and K_s SUs transmitting over a set of N sub-channels, we formulate an optimization problem to find the degree distribution that maximizes the probability of successfully recovering the SUs' signals through the

SIC process, while satisfying the IP constraints of the PUs. The CBS has the perfect knowledge of the PUs channel allocation, power, and respective IP constraints. It also has knowledge of the number of active SUs and their respective SINR constraints. Accordingly, the optimization problem can be formulated as follows:

$$\begin{aligned} &\max_{\mathbf{p},\Omega(x)} && \sum_{k=1}^{K_s} \Pr(S_k) && \\ &\text{s.t.} && \text{(i)} \sum_{i=1}^{d_m} \Omega_i = 1,\ \Omega_i \geq 0, && \forall 1 \leq i \leq d_m \\ & && \text{(ii)} \mathbb{E}\left[\sum_n I_{p,n}^{(k)}\right] \leq I_{\text{th}}^{(k)}, && \forall k \in \mathcal{K}_p. \end{aligned}$$

Condition (i) ensures the sum of all probabilities is equal to 1. Condition (ii) ensures that the PUs are protected by the IP constraints on a per device basis. With reference to Lemma 2, it can easily be seen that this condition determines the value of $\bar{\Omega}$ and $\mathbf{p}$. Optimization is carried out using the Covariance Matrix Adaptation Evolution Strategy (CMA-ES)[96] and can be easily modified for different IP and SINR thresholds.

3.5.4 Practical Considerations for CRNs

In our system, the CBS is assumed to have the perfect knowledge of the PUs' activity and channel conditions, their respective IP constraints, the number of active SUs and their respective SINR threshold. We assume this is made known to the CBS over the control channel, where transmissions are deterministic in duration and nature. Accordingly, the CBS can find the received power constraints necessary and the optimal degree distribution to meet the system constraints. Then, the control channel can also be used to make the degree distribution known to the SUs. As the SUs are assumed to be able to estimate their channel gains from the downlink given the reciprocity of the channel, the signalling overhead is significantly reduced in comparison to fixed resource allocation.

For a given transmission block, the SIC process cannot be initiated without the

knowledge of how many and which sub-channels were accessed by which devices. We assume the SUs share the same seed with the CBS to determine the number and index of the chosen sub-channels through a pre-defined pseudo-random number generator [41].

Finally, it is worthy of noting that the IP constraint can be defined as either the average IP constraint or the peak IP constraint. However, throughout this chapter, we only consider the former definition. This was justified in [97], where it was shown that the average IP constraint provides a higher system capacity than that of the peak IP.

3.6 Numerical Results

3.6.1 General SINR-Based Random Access

Figure 3.6 shows the achievable recovery error probabilities of the SIC process for the clean packet model and the SINR model in an asymptotic setting where the recovery error probability for $C \leq C^*$ is essentially zero. On the other hand, Figure 3.7 shows the achievable recovery error probabilities of both models for N = 200. Simulation results were averaged over a total of 10,000 iterations. As predicted from Proposition 3.2, the SINR model always yields lower or equal recovery error probabilities in comparison to the clean packet model. That is particularly the case for higher system loads. In addition, we see that the lower the SINR threshold, the higher the performance gain in comparison to the clean packet model whose performance is unaffected by the SINR threshold. This can be further justified by considering how the SINR model makes use of all received transmissions over the different sub-channels in each iteration of the SIC process, rather than the clean transmissions only. Finally, we see that the performance of the SINR model for $P_o = \gamma_{\text{th}} = 5$ dB matches that of the clean packet model even for the non-asymptotic case.

It is worthy of noting that the recovery error probabilities found through simulations are higher than that expected from Lemma 1. According to Lemma 1, only the

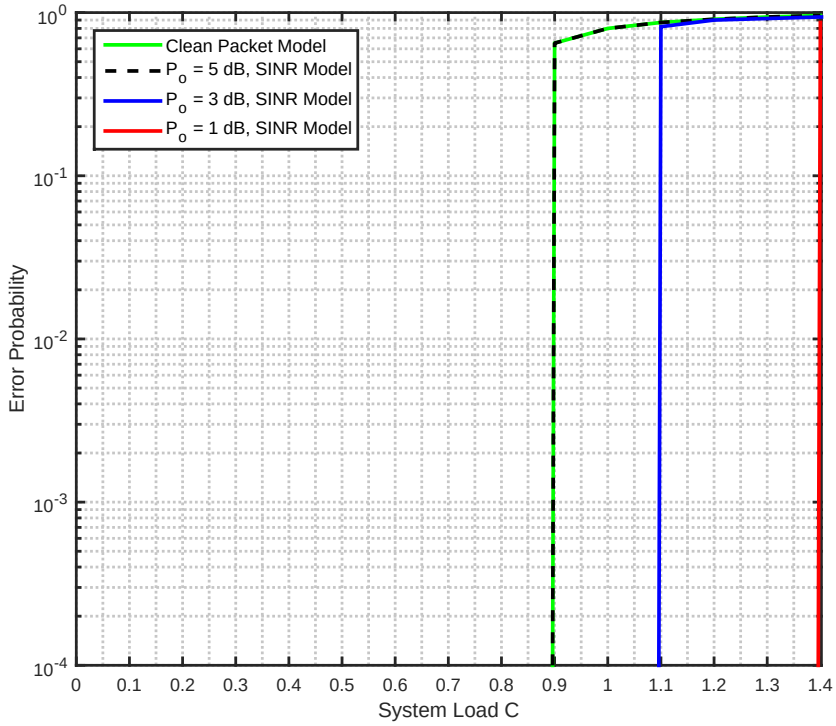

Figure 3.6 – Asymptotic recovery error probability of the clean packet model and the SINR model for $P_o = \gamma_{\text{th}}$.

ratio K/N affects the system performance rather than the actual values of K and N. However, that is provided that $K, N \to \infty$. Thus, we would expect the simulation results to approach more and more that of Lemma 1 as the values of K and N increase.

Finally, it is important to make mention that the average degree $\bar{\Omega}$ directly corresponds to the average received power. Moreover, by averaging over all possible channel realizations, the average received power would also correspond to the average transmit power required by the system. Similarly, the maximum degree d_m corresponds to the peak received power and eventually to the peak transmit power. As shown in Table 3.2, the optimal degree distributions obtained have approximately the same average and peak power requirements. Therefore, we can say that the SINR model is more energy efficient than the clean packet model.

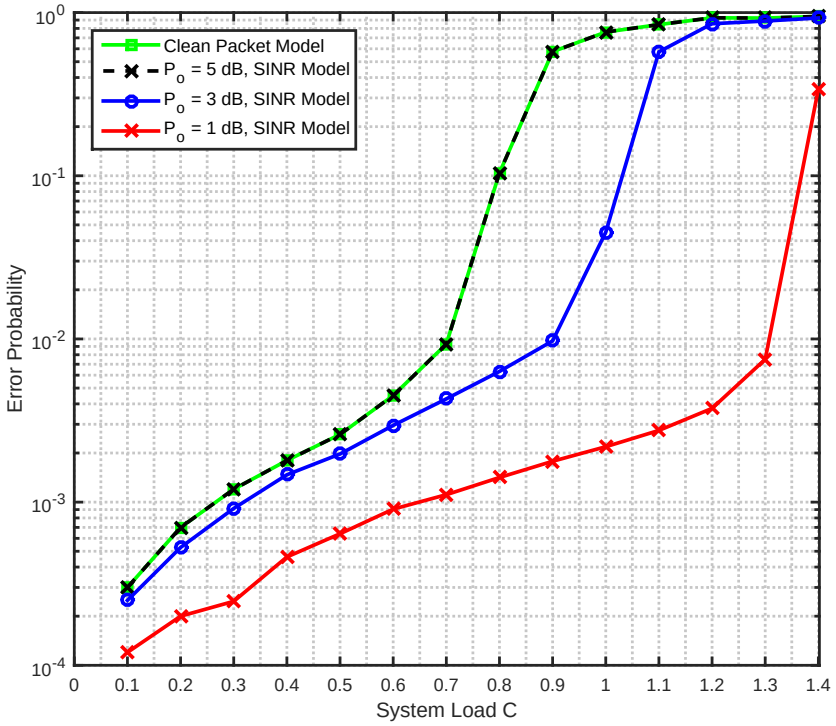

Figure 3.7 – Simulated recovery error probability of the clean packet model and the SINR model for $P_o = \gamma_{\text{th}}$ and $N = 200$.

3.6.2 SINR-Based Random Access in CRNs

We now investigate the system performance in a CRN under different setups. Results are averaged over 10000 iterations. For ease of analysis, we now assume that $P_n = P_o$, where $1 \leq n \leq N$. This condition dictates that all sub-channels have the same received power constraint. For a practical system, this also reduces the signalling overhead. This can be easily justified for the case where $N_p^{(i)} = N_p$ and $I_{\text{th}}^{(i)} = I_{\text{th}}$, for $1 \leq i \leq K_p$. The received power constraint per sub-channel P_o is taken to be 0 dB, and the number of sub-channels N is set to 128 [98]. We adopt this assumption in our simulations. We also assume that $\gamma_{\text{th}}^{(i)} = \gamma_{\text{th}}$, for $1 \leq i \leq K_s$.

In Table 3.4, we show the results of CMA-ES for $\frac{K_s}{N} = \frac{60}{128}$ and a maximum degree of 4 and 8. Using the results of [41], we proceed to compare the achievable error

Table 3.4 – Optimal Degree Distributions for $\frac{K_s}{N} = \frac{60}{128}$

$\log_{10}\gamma_{\text{th}}/P_o$	0 dB		1 dB	
d_m	4	8	4	8
Ω_1	0.0002	0.0003	0.0001	0.0002
Ω_2	0.5072	0.0831	0.1108	0.1295
Ω_3	0.0041	0.1619	0.1727	0.1529
Ω_4	0.4885	0.1744	0.7163	0.2112
Ω_5		0.2255		0.0674
Ω_6		0.0382		0.2406
Ω_7		0.1758		0.1188
Ω_8		0.1408		0.0794
ϵ	3.00e-03	3.75e-04	2.67e-04	4.05e-04

probabilities of both designs; the error probability is denoted by ϵ and defined as $1 - \frac{1}{K_s} \sum_{k-1}^{K_s} \mathbb{P}_{s,k}$. Results are shown in Figure 3.8. As predicted, the proposed design outperforms the 'clean packet' model even for $\frac{\gamma_{\text{th}}}{P_o} = 1$. As our proposed design relies on the overall received SINR, the sum of all individual SINRs, it makes use of all transmissions over the different sub-channels rather than interference-free transmissions only, thus, achieving better performance for the same power requirements.

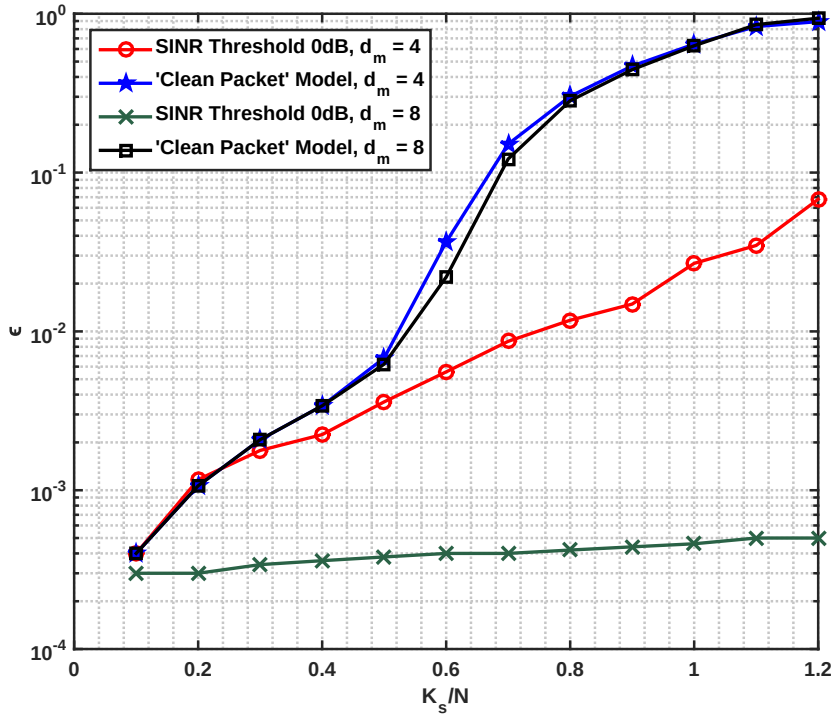

Figure 3.8 – Error probability of proposed scheme in comparison to the 'Clean Packet' model for different ratios of $\frac{K_s}{N}$

Finally, in Figure 3.9, we consider the probability of having the IP caused by the SUs to the PUs below a given threshold. We use the results from Table 3.4, for $d_m = 8$ and $\log_{10} \frac{\gamma_{\text{th}}}{P_o} = 0$dB. We find $\bar{\Omega}$ to be around 5.23. Interestingly, for $I_{\text{th}} \leq$ -5dB, the probability of successfully recovering a PU's signals becomes independent of the threshold and solely dependent on the number of SUs supported in the network. Even more so, for thresholds as high as 10 dB, the probability of successfully recovering a PU's signals is almost one for any number of SUs. It is worthy of noting that

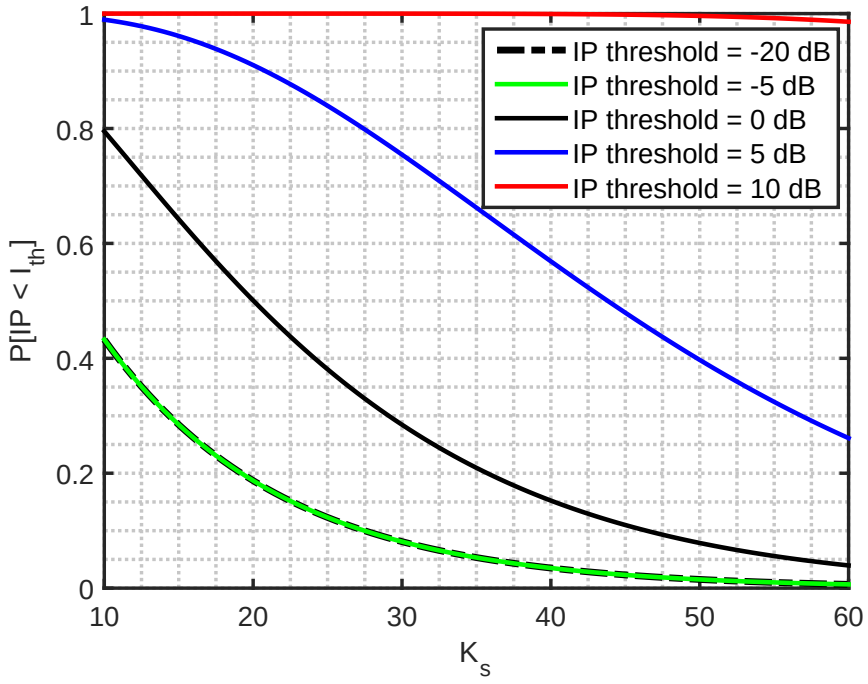

Figure 3.9 – Probability of IP being below the threshold for different values of K_s

Condition (ii) in Section 3.5.3 can be easily modified to limit this probability by ensuring $\Pr\left(\sum_n I_{p,n}^{(k)} \leq I_{\text{th}}^{(k)}\right) \leq \delta$, where δ is a predefined threshold.

3.7 Chapter Summary

In the first part of this chapter, we represented the SIC process in SINR-based RA systems as a novel message passing algorithm. We proposed a novel tree-based analytical framework to track the error probabilities and showed that it is a generalization of the conventional AND-OR tree analysis used for the clean packet model. Finally, we provided analytical and numerical proof that our model achieves higher system loads in comparison to the clean packet model.

In the second part of this chapter, we proposed a new design of probabilistic random access schemes for CRNs. We showed that the conventional 'clean packet' model is sub-optimal under the IP and SINR constraints. Motivated by this, we formulated a

new optimization problem, based on CMA-ES, to maximize the probability of successful recovering the SUs' signals in the SIC process while satisfying the IP constraints of the PUs. Our numerical results showed that our designed degree distributions can achieve lower error probabilities with lower number of transmissions, and thus, with lower power requirements.

Chapter 4

Coded Slotted ALOHA with QoS Guarantees

4.1 Chapter Introduction

4.1.1 Chapter Overview

As mentioned in Chapter 1, one of the main motivations behind our work on the multiple access channel of mMTC is that highly emphasized RACH overload problem in LTE-A [50]. In LTE-A, devices are required to establish an air interface connection prior to data transmission. Access requests are transmitted in an uncoordinated manner over the RACH. Once an MTC device has been granted access, it is scheduled to specific radio resources over which data transmission takes place in a deterministic manner. Readers are referred to Section 2.3 for the details of the process. Nevertheless, with the massive number of MTC devices forecasted to operate in the near future, the RACH in current access schemes will be overloaded and will suffer from continuous collisions requiring multiple re-transmissions. This will result in a large energy expenditure, unexpected delays, and time-frequency resource wastage. Some solutions have been proposed and even standardized to mitigate the RACH overload problem [99, 51], e.g., access class barring schemes, separate RACH resources

for M2M, dynamic allocation of RACH resources, backoff schemes, slotted access, pull-based and group-based schemes.

Aside from the RACH overload problem, establishing energy efficient access schemes with QoS guarantees is another design challenge in M2M networks. We list some of the works proposed in this direction. A group-based approach was proposed in [100] where devices are grouped based on their latency requirements. Each group is allocated specific time intervals with predefined durations over which data transmission takes place in a deterministic manner. The allocated time intervals are non-overlapping and their periodicity is proportional to the packet arrival rates. Thus, groups with higher packet arrival rates are assumed to have tighter latency requirements. An extension of this work can be found in [101] where this assumption is dropped and a generalized access management scheme was proposed with an adaptive resource allocation scheme based on the incoming traffic. In [102], devices with the same QoS requirement are scheduled to transmit over the same resources simultaneously. Using well-designed codes, the superposed codewords can be reliably decoded at the BS allowing for a more efficient use of resources and reduction in the number of retransmissions at the RACH. Other approaches [103, 104, 105] have studied uncoordinated access schemes which do not use any RACHs and instead assign all the resources as uplink data channels. These works have shown potential performance gains in supporting a larger number of devices in comparison to coordinated access schemes for small packet transmissions. In the following, we review the recent works tackling the massive access problem through uncoordinated access which is the main focus of this chapter.

For a single cell scenario with a large number of devices contending to access the same BS, the slotted ALOHA protocol has been widely studied (ref. Section 2.2. Authors in [106] extended this model to a heterogeneous slotted ALOHA setting using the extended AND-OR tree framework derived in [107]. In [106], the heterogeneity in the network represented the different packet loss rates amongst the devices corresponding to their different channel conditions. Using a similar analytical framework, authors in [108] considered the heterogeneous QoS requirements in the network and formulated a multi-edge-type density evolution optimization problem to find the probabilities

and the system load that maximize the overall system utility. This was later shown to be equivalent to single-edge-type density evolution in [109]. Authors in [109] also extended existing finite frame length error floor approximations to the multi-class case and propose a heuristic approach to optimize the degree distributions in the finite frame length regime for low-medium system loads. All these works on heterogeneous coded slotted ALOHA focused on the heterogeneous reliability requirements, e.g., packet loss rate. However, the extension of these transmission schemes to the case of heterogeneous latency requirements is not straightforward. In this chapter, we address this issue by dividing the transmission scheme into multiple stages whose durations are determined by the latency requirements of the different existing classes of devices. Based on this, we develop the analytical expressions and formulate the designs and optimization problems to satisfy the unique requirements and design constraints of M2M communications.

4.1.2 Chapter Contributions

The main contributions of this chapter are summarized below.

Random Acess Scheme

We propose a generalized slotted uncoordinated data transmission scheme for the case of diverse QoS requirements. We consider two different schemes where devices are grouped based on their QoS requirements. The first scheme is called the ACK-All scheme. In this scheme, MTC devices from all groups transmit simultaneously over the same radio resources in all stages of the transmission frame. The second scheme is called the ACK-Group scheme. In this scheme, MTC devices from different groups transmit in distinct stages. We discuss the advantages of these schemes over coordinated access schemes. We show that the proposed RA schemes can service a larger number of devices when the number of time-frequency resources is sufficiently large, while guaranteeing the diverse QoS requirements. We further show that the ACK-All scheme is advantageous over the ACK-Group scheme when the number of

devices with tighter latency requirements is less than that with more flexible latency requirements.

AND-OR Tree Based Performance Analysis

We use the AND-OR tree to analyze the system performance characterized by the average probability of device resolution error for both transmission schemes: ACK-All and ACK-Group. The derived analytical expressions can be seen as a generalization of those in [108] which only characterizes the error probabilities at the end of the first sub-frame. As the devices that are acknowledged do not retransmit in following sub-frames, the intermediate feedback introduced between the sub-frames results in graphs with reduced sets of nodes as well as reduced edges. We characterize these reductions by reformulating the AND-OR tree expressions based on the reduced sets of active devices and slots along with their reduced degree distributions[1]. We validate the accuracy of the expressions under different settings using simulations.

System Design for QoS Guarantees

We show how the derived expressions can be used to design systems that can guarantee the QoS requirements of different groups with significantly high energy efficiency and high reliability for the proposed scheme. For this, we cannot use the optimization problem formulated in [108]. Authors in [108] simplify their optimization problem by assuming vanishing error probabilities, i.e., all classes have a packet loss rate of zero. However, the error probabilities cannot be assumed to be vanishing at the end of each sub-frame in the proposed scheme. For example, lower priority groups can tolerate larger delays and, thus, their error probabilities will be far from vanishing in the early stages of transmission. Moreover, vanishing error probabilities are only relevant to ultra-reliable applications [110] which are only a subset of M2M applications. Therefore, we need to guarantee the diverse latency requirements with diverse error

[1]It is worthy of noting that the work in this chapter can be directly extended to the SINR-based model in the previous chapter at the expense of a slight increase in analytical complexity.

probabilities. Accordingly, we propose a guideline that allows us to design the access probabilities using the AND-OR tree, which was originally designed for asymptotically large systems, for a finite number of devices and resources.

4.1.3 Chapter Outline

The rest of this chapter is organized as follows. In Section 4.2, we present the system model and the proposed RA schemes. In Section 4.3, we consider a tree-based analytical framework and derive the expressions for the average probabilities of device resolution. An energy efficient system design is discussed in Section 4.4 along with the limitations on the system load. A reliable system design is discussed in Section 4.5 and is shown to be valid for a finite number of devices. Numerical results and practical considerations are presented in Section 4.6. Conclusions are drawn in Section 4.7.

The notations used in this chapter are summarized in Table 4.1 for quick reference.

4.2 System Model

4.2.1 Overview

We consider a scenario where K uniformly distributed MTC devices communicate with a BS located at the origin. The devices are assumed to have fixed locations. We consider the uplink of an OFDMA system, where the frequency resources are divided into several sub-channels each with a bandwidth Δf. A radio resource unit consists of a sub-channel along with a time slot of duration T. We characterize the QoS of MTC devices in terms of the delay as diverse M2M applications have diverse latency requirements. More specifically, we consider a batch arrival model with different delay groups denoted by $\mathcal{C}_1, \mathcal{C}_2, ..., \mathcal{C}_r$. Their respective delay requirements are quantified in terms of time slots and denoted by $N_1, N_2, \ldots, N_r$, where $N_i < N_{i'}$ for $i < i'$, i.e.,

Table 4.1 – Notation Summary

Notation	Description
K	Number of MTC devices
r	Number of groups of MTC devices
N_i	Latency requirement of the i^{th} group of MTC Devices
$\mathcal{C}_i$	Group of MTC devices with a latency requirement of N_i
$\mathcal{C}_{i,k}^{(s)}$	Subset of MTC devices in $\mathcal{C}_i$ that potentially transmitted in the first k sub-frames, $k \leq s$
ΔN_i	Number of times slots in sub-frame i
N	Total number of time slots in a frame
$p_i^{(s)}$	Access probability allocated to $\mathcal{C}_i$ in sub-frame s
$g_i^{(s)}$	Average degree of a check node in sub-frame s
$\zeta_i^{(s)}$	Average degree of a variable node in $\mathcal{C}_{i,s}^{(s)}$
$\epsilon_i^{(s)}$	Average probability of device resolution error of $\mathcal{C}_i$ after s sub-frames
ϵ_i^*	Target probability of device resolution error of $\mathcal{C}_i$
α_i	Fraction of MTC devices in $\mathcal{C}_i$
β_s	Fraction of time slots in sub-frame s
M_i	Average number of transmissions of a device in $\mathcal{C}_i$
γ_{ref}	Received SNR of a device's packet

devices from the group $\mathcal{C}_i$ have a tighter latency requirement than devices from the group $\mathcal{C}_{i'}$ for $i < i'$.

The resources allocated for M2M may either be fixed or may vary from one frame to the other. For every transmission frame, the BS is assumed to be capable of estimating the number of active devices and grouping them based on their QoS requirements. Similarly, the number of active devices and the number of groups may either be fixed or may vary from one frame to the other. Active devices are devices with at least one packet to transmit. Furthermore, we assume devices transmit one packet within one transmission frame. Packets are assumed to be of equal size and can be transmitted in one radio resource unit.

We assume a block fading channel, i.e., the channel gains remain unchanged for the duration of the transmission frame and vary randomly and independently from one frame to the other. Channels are also assumed to be reciprocal, i.e., uplink and downlink channel states are the same. Therefore, each device can estimate the uplink

channel gain from the pilot signal sent periodically over the downlink channel by the BS. The devices also use these pilot signals to synchronize their timing to that of the BS [111, 112]. The devices perform channel inversion such that they all have the same received SNR γ_{ref} [101]. γ_{ref} is determined by the coding and modulation schemes adopted in the system such that the BS can reliably recover a device's packet if its SNR is greater than or equal to γ_{ref}. In reality, MTC devices are power limited. For channels with low SNR, the device may not always be able to perform channel inversion. In this case, the device's channel is said to be in outage and the device remains silent.

4.2.2 Probabilistic Data Transmission

Each transmission frame is of length $N = N_r$ time slots. The frame is divided into r sub-frames, where the length of sub-frame s is denoted by ΔN_s and is equal to $N_s - N_{s-1}$, for $1 \leq s \leq r$ and $N_0 = 0$. For a given sub-frame s, we consider the collision channel model and transmission scheme in [46, 47, 48, 107]. The BS assigns every group $\mathcal{C}_i$ an access probability $p_i^{(s)}$. That is, a device in group $\mathcal{C}_i$ transmits in a given time slot of sub-frame s with a probability $p_i^{(s)}$ and remains silent with a probability $1 - p_i^{(s)}$. We emphasize that the number of probabilities that are to be distributed to the devices is equal to the number of different groups and not the number of devices. Therefore, the signalling overhead is fairly small and can be easily incorporated alongside the pilot signals, beacons, acknowledgements and other control data.

The BS is assumed to be able to distinguish between idle slots, singleton slots, and collision slots. This is feasible with the assumed power control strategy. Once the BS detects the reception of a singleton slot, the transmitted packet is recovered and the respective device is said to be resolved. The BS and the devices can share a common pseudo-random generator function that computes a random seed based on the device identity. The devices use these seeds to generate the indices of the slots they want to transmit in. Once a device's packet is recovered at the BS, the BS can extract the

respective device's identity from the header file. Then, the BS can generate the seed and the necessary indices of slots to perform SIC. That is, the copies of the recovered packet are cancelled from the remaining slots. This allows for the potential of having more singleton slots in the following iteration, and, thus, the potential of recovering more packets. We assume perfect SIC.

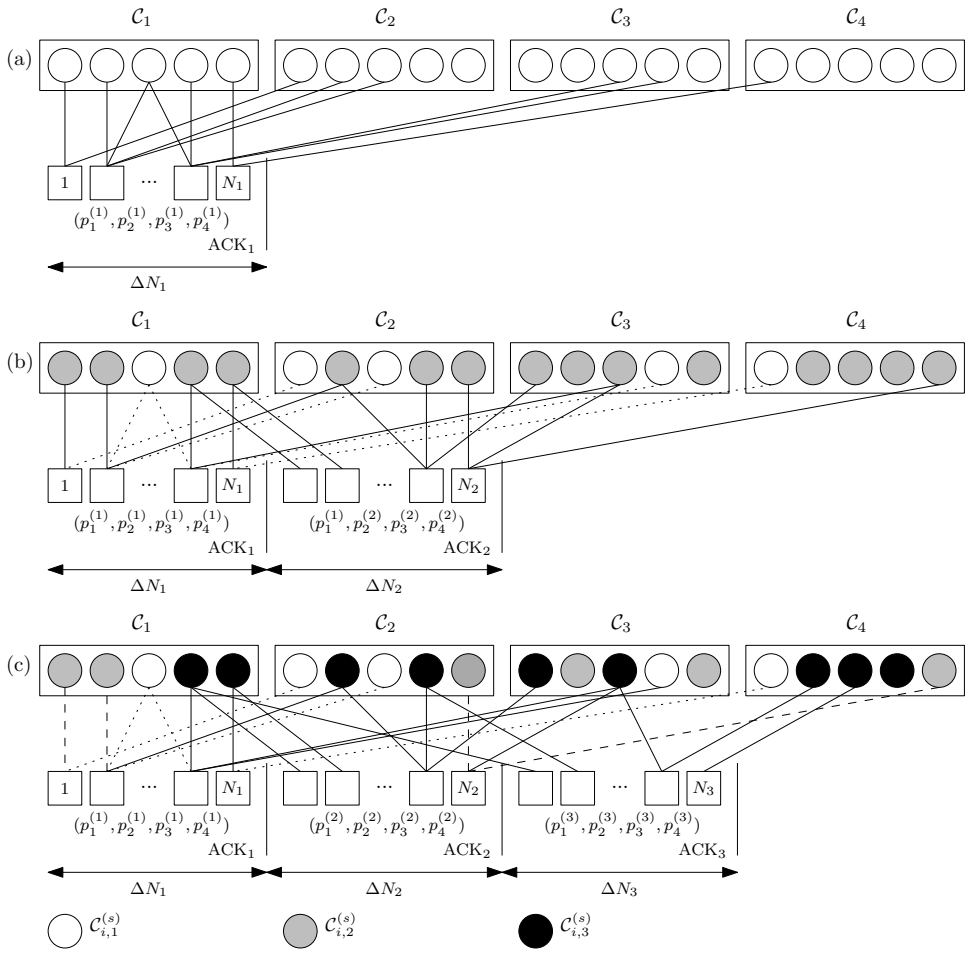

Figure 4.1 – The ACK-All scheme (r = 4) in three consecutive sub-frames. The dotted and dashed lines denote the edges that were removed at the receiver side after the first and second sub-frame respectively, and the solid lines denote the edges that have remained.

The BS performs SIC at the end of each sub-frame s using all the received packets over the N_s time slots. We assume a perfect feedback channel. A batch of acknowledgements is sent at the end of each sub-frame to all successfully resolved devices,

and all acknowledged devices do not transmit in future time slots. We refer to this scheme as the ACK-All scheme. The transmission scheme in [108] can be seen as a special case with a single sub-frame of our more generalized ACK-All scheme. For a heterogeneous random access setting, having devices from different groups share the same resources and transmit simultaneously over the same frames results in a performance degradation to the lower-prioritized devices [108]. This behavior was originally noted for codes-on-graph in [113]. In [C3]-[C2], we showed that by dividing the transmission of codes-on-graph into multiple stages separated by intermediate feedback, we can better serve these lower-prioritized devices while still guaranteeing the diverse QoS requirements. In this chapter, we extend this concept to the random access setting.

An example of the ACK-All scheme is illustrated in Figure 4.1 in terms of a bipartite graph. The circles represent the devices and are referred to as Variable Nodes (VNs). The squares represent the time slots and are referred to as Check Nodes (CNs). An edge between two nodes indicates that the device transmitted in the respective time slot. No edge between two nodes indicates that the device was silent in the respective time slot. As mentioned before, the BS performs SIC by cancelling the copies of the recovered packet from the remaining slots. In that case, the edges connecting the resolved VNs to the CNs are removed from the graph. Moreover, we denote by $\mathcal{C}_{i,k}^{(s)}$ the subset of devices in group $\mathcal{C}_i$ that potentially transmitted in the first k sub-frames, where $1 \leq k \leq s$.

We consider another scheme, namely, the ACK-Group scheme. In this case, devices from the group $\mathcal{C}_i$ are scheduled to transmit in sub-frame i only and remain silent otherwise. The two schemes are illustrated and compared in Figure 4.2. For the ACK-Group scheme, transmissions of devices from different groups are shown to be separated. It is worth mentioning that Figure 4.1 assumes flexible latency requirements as unresolved devices continue to transmit even after their delay has been violated, whereas Figure 4.2 assumes strict latency requirements, i.e., unresolved devices are dropped from the network once their latency requirement is violated.

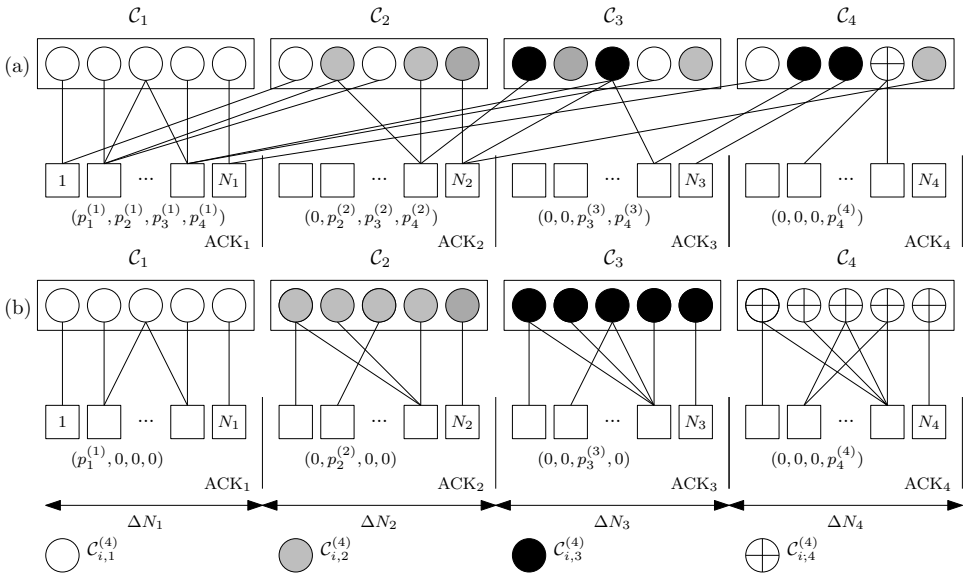

Figure 4.2 – Overall bipartite graph representation of the two transmission schemes with $r = 4$: (a) ACK-All and (b) ACK-Group.

4.3 AND-OR Analysis of the Proposed RA Schemes

The number of edges branching out of a node is said to be the degree of the respective node. These degrees are random variables that play an important role in the system performance. In this section, we first derive the expressions of the degree distributions of the VNs and the CNs. We also introduce the AND-OR tree which is a well-known tool for calculating the decoding error probabilities of information symbols in codes on graph. Then, we propose a more generalized AND-OR tree to model the system with the ACK-All and ACK-Group schemes.

4.3.1 Degree Distributions

For generality purposes, we assume flexible latency requirements, i.e., $p_i^{(s)} \geq 0$ for $i > s$. We denote by $\mathcal{G}_s$ the bipartite graph formed by the first s sub-frames. As mentioned before, for a given sub-frame s, each group $\mathcal{C}_i$ can be divided into s subsets: $\mathcal{C}_{i,1}^{(s)}, \mathcal{C}_{i,2}^{(s)}, \ldots, \mathcal{C}_{i,s}^{(s)}$.

The probability that a device from $\mathcal{C}_{i,s}^{(s)}$ transmits in a given time slot of sub-frame s

is a Bernoulli random variable with a success probability of $p_i^{(s)}$. As this probability is independent of previous transmissions, the probability that a device from the group $\mathcal{C}_{i,s}^{(s)}$ transmits in d time slots in a sub-frame s is the sum of d i.i.d. Bernoulli random variables and is denoted by $\Lambda_{i,d}^{(s)}$. From [46, 47, 48, 107], this is equivalent to the following binomial distribution:

$$\Lambda_{i,d}^{(s)} = \binom{\Delta N_s}{d} \left(p_i^{(s)}\right)^d \left(1 - p_i^{(s)}\right)^{\Delta N_s - d}. \tag{4.1}$$

Furthermore, the probability that d devices from the group $\mathcal{C}_{i,s}^{(s)}$ transmit in a given time slot of sub-frame s, is also the sum of d i.i.d. Bernoulli random variables. From [46, 47, 48, 107], the equivalent binomial distribution is given as

$$\Omega_{i,d}^{(s)} = \binom{|\mathcal{C}_{i,s}^{(s)}|}{d} \left(p_i^{(s)}\right)^d \left(1 - p_i^{(s)}\right)^{|\mathcal{C}_i^{(s)}| - d}. \tag{4.2}$$

For convenience, let $p_i^{(s)} = \frac{g_i^{(s)}}{|\mathcal{C}_{i,s}^{(s)}|}$, where $0 \leq g_i^{(s)} \leq |\mathcal{C}_{i,s}^{(s)}|$ is the average number of devices from the group $\mathcal{C}_{i,s}^{(s)}$ that access a given time slot of sub-frame s. For sufficiently large K and N, the expression in Equation 4.1 can be approximated, as in[46, 47, 48, 107], by the following Poisson distribution

$$\Lambda_{i,d}^{(s)} = \frac{\left(\zeta_i^{(s)}\right)^d \exp\left(-\zeta_i^{(s)}\right)}{d!}, \quad \text{for } 1 \leq k \leq s, \tag{4.3}$$

where $\zeta_i^{(s)} = \frac{g_i^{(s)} \Delta N_s}{|\mathcal{C}_{i,s}^{(s)}|}$ is the average number of time slots from sub-frame s that are selected by a given device from the group $\mathcal{C}_{i,s}^{(s)}$. Similarly, the expression in Equation 4.2 can be approximated by the following

$$\Omega_{i,d}^{(s)} = \frac{\left(g_i^{(s)}\right)^d \exp\left(-g_i^{(s)}\right)}{d!}. \tag{4.4}$$

We now consider the generator polynomial $\Psi_i^{(s)}(x) = \sum_d \Psi_{i,d}^{(s)} x^d$ to represent the

overall degree distribution of a VN in group $\mathcal{C}_{i,s}^{(s)}$ in the bipartite graph $\mathcal{G}_s$. With a slight abuse of notation, the superscript here denotes the sum degree over the first s sub-frames rather than that of a single sub-frame s as in Equation 4.1. For every VN, we define a vector $\mathbf{m}$ of dimension s, where its i^{th} element, m_i, corresponds to its number of transmissions in the i^{th} sub-frame. m_i is an independent binomial random variable with a distinct success probability as given in Equation 4.1. The overall degree of a VN is the sum of all elements of $\mathbf{m}$ and, thus, follows a Poisson binomial distribution. We also define $\mathcal{M}_j^{(s)} \triangleq \{\mathbf{m} : \sum_{k=1}^{s} m_k = j\}$, for $1 \leq j \leq N_s$. The set $\mathcal{M}_j^{(s)}$ contains all possible realizations of $\mathbf{m}$ whose elements add up to an overall degree of j. Therefore, in the subgraph $\mathcal{G}_s$, the probability that a VN of group $\mathcal{C}_{i,s}^{(s)}$ is of degree j can be expressed as

$$\Psi_{i,j}^{(s)} = \sum_{\mathbf{m}\in\mathcal{M}_j^{(s)}} \prod_{s'=1}^{s} \Lambda_{i,m_{s'}}^{(s')}. \tag{4.5}$$

We also consider the generator polynomial $\Delta^{(s)}(x) = \sum_d \Delta_d^{(s)} x^d$ to represent the overall degree distribution of a CN in sub-frame s. For every CN, we define a vector $\mathbf{l}$ of dimension r, where its i^{th} element, l_i, corresponds to the number of transmissions from the i^{th} group, $\mathcal{C}_i$. l_i is an independent binomial random variable with a distinct success probability as given in Equation 4.2. Thus, the overall degree also follows a Poisson binomial distribution. We also define $\mathcal{L}_j^{(s)} \triangleq \{\mathbf{l} : \sum_{i=1}^{r} l_i = j\}$, for $1 \leq j \leq \sum_{i=1}^{r} |\mathcal{C}_{i,s}^{(s)}|$. $\mathcal{L}_j^{(s)}$ contains all possible realizations of $\mathbf{l}$ whose elements add up to an overall degree of j. Then, we can write

$$\Delta_j^{(s)} = \sum_{\mathbf{l}\in\mathcal{L}_j^{(s)}} \prod_{i=1}^{r} \Omega_{i,l_i}^{(s)}, \quad \text{for } 1 \leq l \leq r. \tag{4.6}$$

For sufficiently large K and N, the sum degree can be seen as the sum of independent Poisson random variables with different averages, which is also a Poisson random variable with an average equal to the sum of its individual components. Accordingly,

we can rewrite Equation 4.5 as

$$\Psi_i^{(s)}(x) = \exp\left(-\sum_{s'=1}^{s} \zeta_i^{(s')}(1-x)\right). \tag{4.7}$$

Based on the same argument, we can rewrite Equation 4.6 as

$$\Delta^{(s)}(x) = \exp\left(-\sum_{i=1}^{r} g_i^{(s)}(1-x)\right). \tag{4.8}$$

From an edge perspective, the generator polynomials corresponding to the degree distributions of both VNs and CNs are defined as follows:

$$\psi_i^{(s)}(x) \triangleq \frac{\sum_d d\Psi_{i,d}^{(s)} x^{d-1}}{\bar{\Psi}_i^{(s)}} \text{ and } \delta^{(s)}(x) \triangleq \frac{\sum_d d\Delta_d^{(s)} x^{d-1}}{\bar{\Delta}^{(s)}}, \tag{4.9}$$

where $\bar{\Psi}_i^{(s)} = \sum_d d\Psi_{i,d}^{(s)}$ and $\bar{\Delta}^{(s)} = \sum_d d\Delta_d^{(s)}$ are the average degrees of the devices and the slots in sub-frame s, respectively.

4.3.2 Tree Assumption

Consider an edge (v, c) chosen uniformly at random from $\mathcal{G}_s$ connecting a VN v to an arbitrary CN c. By removing that edge, a subgraph is generated by v and all the neighbors of v within distance 2ℓ. This subgraph was shown in [45] to be a tree asymptotically with v being its root at depth 0 and its leaves at depth 2ℓ. Nodes at depth i have children at depths $i+1$. The CNs are located at depth $1, 3, ..., 2\ell - 1$, and the VNs are located at depth $0, 2, ..., 2\ell$. For a given iteration of the SIC process, an edge of a CN can be removed if and only if all remaining edges have been removed in previous iterations. Thus, CNs act as AND-nodes. Conversely, an edge of a VN can be removed if at least one of the remaining edges has been removed in previous iterations. Thus, VNs act as OR-nodes. Hence these trees are often referred to as AND-OR trees [45].

AND-nodes are categorized into s different types where each type corresponds to one of the sub-frames. Similarly, OR-nodes are categorized into r different types where

each type corresponds to one group of devices. For each graph $\mathcal{G}_s$, we consider r trees where the i^{th} tree is denoted by $T_{i,\ell}^{(s)}$ with depth 2ℓ and a Type-i OR-node at its root (depth 0). Initially, each Type-i OR-node at depth 2ℓ is assigned a value 0 with a probability $q_i^{(s)}[0]$ and is 1 otherwise. We are interested in finding the probability that the root of each tree evaluates to 0.

In general, the tree assumption holds with high probability provided that the frame length is large enough [114, 42, 115]. For lack of space, we do not derive the explicit requirements for this assumption to hold. Instead, we rely on numerical evidence to prove the accuracy of the tree- assumption in characterizing the system performance in our case.

4.3.3 ACK-All Transmission Scheme

We define $\epsilon_i^{(s)}$ as the average fraction of devices in $\mathcal{C}_i$ that remain unresolved at the end of sub-frame s. The following lemma models the evolution of the average error probabilities in a given SIC iteration at the end of the transmission sub-frame s for the ACK-All scheme (Figure 4.2) and calculates $\epsilon_i^{(s)}$ for $1 \leq i, s \leq r$.

In the first sub-frame, all the packets are unknown and, thus, the expressions are simplified to those in [108]. However, as the devices that are acknowledged do not retransmit in following sub-frames, the intermediate feedback introduced at intermediate stages of the transmission frame will lead to reductions in the graph that need to be carefully characterized. Namely, in the following lemma, we characterize the reduced sets of active devices and slots along with their reduced degree distributions to derive the expressions for the error probabilities for each group at the end of every sub-frame. Readers are referred to Section B.1 for the proof of this lemma.

Lemma 3. *For the ACK-All scheme, the probability that a device from the group* $\mathcal{C}_i$

remains unresolved at the end of sub-frame s is given below as:

$$\epsilon_i^{(s)} = \lim_{\ell\to\infty} q_i^{(s)}[\ell], \quad \textit{where}$$
$$q_i^{(s)}[\ell] =$$
$$\psi_i^{(s)}\left(1-\sum_{s'=1}^{s}\bar{c}_i^{(s')}\delta^{(s')}\left(1-\sum_{i'=1}^{r}\bar{v}_{i'}^{(s')}\frac{q_{i'}^{(s)}[\ell-1]}{q_{i'}^{(s')}[0]}\right)\right),$$
$$q_i^{(s)}[0] = \epsilon_i^{(s-1)}, \quad q_i^{(1)}[0] = 1,$$
$$\bar{v}_i^{(s)} = \frac{p_i^{(s)}|\mathcal{C}_{i,s}^{(s)}|}{\sum_{i'=1}^{r} p_{i'}^{(s)}|\mathcal{C}_{i',s}^{(s)}|} = \frac{g_i^{(s)}}{\sum_{i'=1}^{r} g_{i'}^{(s)}}, \ \textit{and}$$
$$\bar{c}_i^{(s)} = \frac{p_i^{(s)}\Delta N_s}{\sum_{s'=1}^{s} p_i^{(s')}\Delta N_{s'}} = \frac{\zeta_i^{(s)}}{\sum_{s'=1}^{s}\zeta_i^{(s')}}.$$

4.3.4 ACK-Group Transmission Scheme

In the case of separate transmissions, the transmissions of devices within a group are non-overlapping with the transmissions of devices from other groups. Therefore, the CNs in each sub-frame are only connected to one group of VNs. In other words, Type-i AND-nodes are only connected to Type-i OR-nodes, for $1 \le i \le r$. Following on the proof of Lemma 3, the ACK-Group scheme is a special case of our proposed scheme where $g_i^{(s)} = 0$ for $i \neq s$. The following proposition is then derived.

Proposition 4.1. *For the ACK-Group scheme, the probability that a device from the group $\mathcal{C}_i$ remains unresolved at the end of sub-frame i is given below as:*

$$\epsilon_i^{(i)} = \lim_{\ell\to\infty} q_i[\ell], \ \textit{where}$$
$$q_i[\ell] = \lambda_i^{(i)}\left(1-\omega_i^{(i)}\left(1-q_i[\ell-1]\right)\right), q_i[0] = 1,$$
$$\lambda_i^{(s)}(x) \triangleq \frac{1}{\sum_d d\Lambda_{i,d}^{(s)}}\sum_d d\Lambda_{i,d}^{(s)}x^{d-1} \ \textit{and}$$
$$\omega^{(s)}(x) \triangleq \frac{1}{\sum_d d\Omega_{i,d}^{(s)}}\sum_d d\Omega_d^{(s)}x^{d-1}.$$

Another possible data transmission scheme is to acknowledge the resolved devices from $\mathcal{C}_i$ only at the end of sub-frame i. Based on the work in [116] on codes-on-graph, authors showed that the encoding of already decoded information symbols is useful in the evolution of the decoding process at the receiver. However, we found that the transmissions of resolved devices in the following sub-frames were unnecessary and actually degraded the performance of the SIC process at the BS. The scheme performed poorly in comparison to the ACK-All scheme and demonstrated little performance gains in comparison to the ACK-Group scheme [C3]; therefore, we will not consider it in this thesis.

4.4 Design of Energy Efficient RA Schemes

4.4.1 Performance Metrics

For ease of notation, we denote the size of each group $\mathcal{C}_i$ by $\alpha_i K$ where $0 < \alpha_i < 1$ and $\sum_i \alpha_i = 1$. We also define a set of ratios $\beta_1, \beta_2, ..., \beta_r$, where $\beta_i \triangleq \frac{N_i}{N}$. We now list the main performance metrics to be considered:

System Load

The system load is denoted by L and is defined as the ratio of the number of active devices K to the number of time slots in the transmission frame N.

Device Resolution Error

We define another QoS requirement, namely, the target average probability of device resolution error ϵ_i^*, for $1 \leq i \leq r$. This is the maximum acceptable fraction of devices from group $\mathcal{C}_i$ that violate their latency requirement, on average. That is, we need to ensure that $\epsilon_i^{(i)} \leq \epsilon_i^*$.

Blocking Probability

In some cases, the system load may be very large such that there is no solution that can satisfy the QoS requirements, i.e., delay and probability of device resolution. The system is said to be overloaded. In this case, the BS enforces some access barring techniques [99] to block some devices from accessing the network so as not to jeopardize the system performance. Details on this will be explained in Section 4.4.4.

Average Number of Transmissions

M2M devices are notorious for having low power budgets. The average number of transmissions required for a device to meet its QoS requirement should be limited below a certain threshold for an energy efficient system. The average number of transmissions for each group can be calculated from the following proposition. Readers are referred to Section B.2 for the proof of this proposition.

Proposition 4.2. *The average number of transmissions of a device from the group $\mathcal{C}_i$ can be calculated as*

$$M_i = \sum_{s=1}^{r} g_i^{(s)} \frac{\beta_s}{\alpha_i} \frac{N}{K}.$$

4.4.2 Design Objectives

We denote by $\mathbf{G}$ an $r \times r$ matrix given as

$$\mathbf{G} = \begin{pmatrix} g_1^{(1)} & \dots & g_r^{(1)} \\ & \ddots & \\ g_1^{(r)} & \dots & g_r^{(r)} \end{pmatrix}.$$

The main design objective is to find a matrix $\mathbf{G}$ that satisfies the latency requirements with an acceptable average probability of device resolution error. In all what follows, for simplicity, we only consider strict latency requirements, i.e., $g_i^{(s)} = 0$ for $i > s$, which is a special case of the previously derived expressions. Therefore, we have to

solve for $\sum_{i=1}^{r} r - i + 1$ variables instead of r^2 variables. The main design objective becomes to find a matrix $\mathbf{G}$ that satisfies the following constraint:

$$\epsilon_i^{(i)} \leq \epsilon_i^*, \quad \text{for } 1 \leq i \leq r. \tag{4.10}$$

Let $L^*(\epsilon)$ denote the maximum load for which a group of devices can be resolved within the required time with an average error probability of ϵ. The load of each group in each sub-frame is defined as the ratio of the number of unresolved devices in that group to the number of time slots in that sub-frame. Based on this definition, we present the following proposition with the cases where it is possible to satisfy the constraints in Equation 4.10. Readers are referred to Section B.3 for the proof of this proposition.

Proposition 4.3. *There exists a matrix $\boldsymbol{G}$ that can satisfy the latency requirements of all the groups in the ACK-All and ACK-Group schemes if and only if the following condition is satisfied:*

$$\frac{\epsilon_i^{(i-1)}\alpha_i}{\beta_i}\frac{K}{N} \leq L^*\left(\frac{\epsilon_i^*}{\epsilon_i^{(i-1)}}\right), \tag{4.11}$$

where $\epsilon_i^{(0)} = 1$ and $1 \leq i \leq r$.

For the special case when the actual load of a particular group is approximately equal to the bound in Equation 4.11, the devices of the respective group have to transmit separately to guarantee their QoS requirements in the two schemes: ACK-All and ACK-Group. Furthermore, when the actual load of all groups is approximately equal to L^*, the matrix $\mathbf{G}$ that best satisfies the necessary constraints is the same for both schemes. In other words, the best performance is achieved with separate transmissions.

For the ACK-Group scheme, with reference to Proposition 4.3, we have to ensure that $L^*(\epsilon_i^*) \geq \frac{\alpha_i K}{\beta_i N}$. Consider the special case when $\epsilon_i^* = \epsilon$ for $1 \leq i \leq r$. We have the

following design constraint

$$\frac{K}{N} \leq L^*(\epsilon) \min\{\frac{\beta_1}{\alpha_1}, \ldots, \frac{\beta_r}{\alpha_r}\}. \tag{4.12}$$

The arguments of the min function imply that some sub-frames may have a lower system load than others. Consider the case when the number of devices in group $\mathcal{C}_1$ is relatively small in comparison to its latency requirement. Then, it is likely that devices of the first group will be resolved with less than $\beta_1 N$ time slots. In fact, from the definition of $L^*(\epsilon)$, these devices can satisfy their QoS requirement with only $\frac{\alpha_1 K}{L(\epsilon_1^*)}$ time slots. Therefore, for fair comparison, the length of every sub-frame $s < r$ is set to a length of min $\left(\beta_s N, \frac{\alpha_s K}{L(\epsilon_s^*)}\right)$ which will allow the remaining groups more time slots to meet or improve their performance, if needed.

4.4.3 Special case of Two Groups

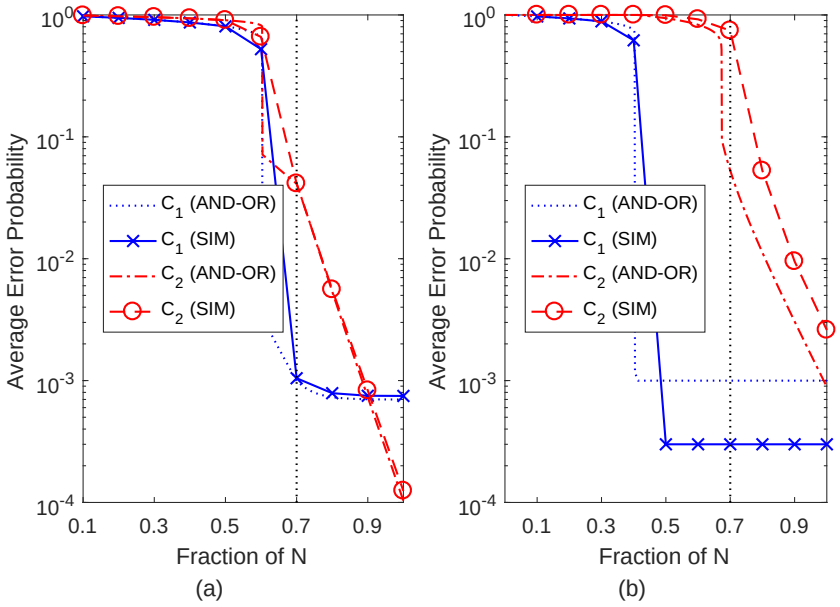

Figure 4.3 – Achievable probabilities of device resolution error for two groups of MTC devices with $N/K = 2.0$, $\beta_1 = 0.7$, $\alpha_1 = 0.5$ and $N = 8000$: (a) ACK-All and (b) ACK-Group.

We compare the two transmission schemes for the same system loads and QoS re-

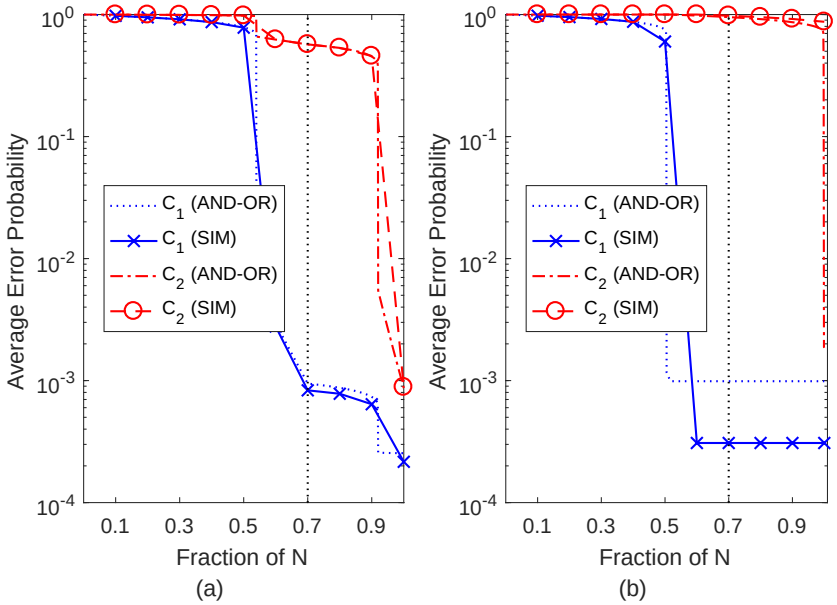

Figure 4.4 – Achievable probabilities of device resolution error for two groups of MTC devices with $N/K = 1.6$, $\beta_1 = 0.7$, $\alpha_1 = 0.5$ and $N = 8000$: (a) ACK-All and (b) ACK-Group.

quirements. Consider the case where $r = 2$. We set $\epsilon_1^* = \epsilon_2^* = 10^{-3}$ and use differential evolution [93] to find $\mathbf{G}$ that minimizes the sum of average transmissions $M_1 + M_2$ whose expressions were given in Proposition 4.2. Figure 4.3 shows the analytical results (AND-OR) for both schemes as well as the simulation results (sim) for $K = 4000$. The simulation results closely match those calculated from the AND-OR tree. For the ACK-All scheme, we can see that group $\mathcal{C}_1$ achieves the desired error probability of 10^{-3} after βN time slots. As $\mathcal{C}_2$ participates in encoding in this sub-frame, $\mathcal{C}_1$ needs more time to reach its target error probability in comparison to the ACK-Group scheme where only devices of $\mathcal{C}_1$ participate in encoding. However, we see the disadvantage of ACK-Group in Figure 4.4. The achievable error probabilities are plotted for a larger system load. In this figure, we can see that even though group $\mathcal{C}_1$ maintains the same performance, group $\mathcal{C}_2$ fails to satisfy its QoS with the ACK-Group scheme. Thus, we say that the ACK-All scheme can service more devices in this case. We will elaborate on this in Section 4.5.

The other disadvantage of the ACK-Group scheme is shown in Figure 4.5a. Although

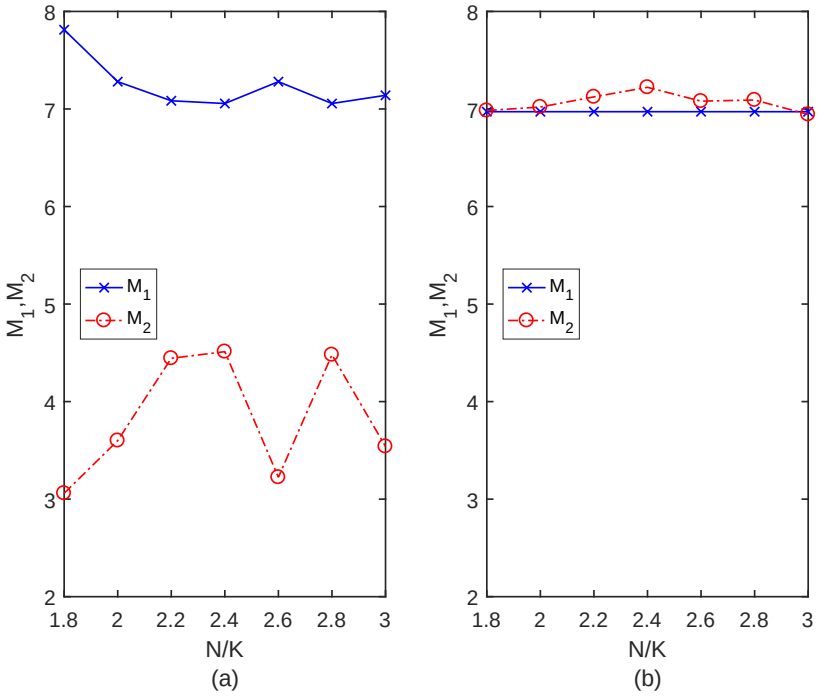

Figure 4.5 – Average number of transmissions of each group of MTC devices with $\beta_1 = 0.7$ and $\alpha_1 = 0.5$: (a) ACK-All and (b) ACK-Group.

M_1 is the same in both schemes, M_2 is reduced in the ACK-All scheme, leading to a reduction in the overall average number of transmissions of all devices in the system as shown in Figure 4.5b. As the average number of transmissions is related to the energy expenditure, we say that the ACK-All scheme is more energy efficient in this case. Nevertheless, we can see that the average number of transmissions for both RA schemes is quite comparable to the number of transmissions allowed in cellular access networks over the RACH [102], which further supports RA as a good candidate for future M2M communication networks. Finally, we would like to note that error floors are a common feature of iterative decoding whether for random access or for codes on graph. Error floors are due to having a non-zero probability of devices not transmitting a packet at all in a frame and, thus, no chance of being recovered at the BS [46]. However, we will show in Section 4.5 that the probability of a device not transmitting is very small. In Section 4.6.1, we will show that this probability is further reduced when unsuccessful devices are allowed to retransmit in following

frames.

4.4.4 Access Barring

Finally, let $L_r^*(\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\epsilon}^*)$ be the maximum system load that the BS can service for the parameters $\boldsymbol{\alpha} = [\alpha_1, \alpha_2, ..., \alpha_r]$, $\boldsymbol{\beta} = [\beta_1, \beta_2, ..., 1]$ and constraints $\boldsymbol{\epsilon}^* = [\epsilon_1^*, \epsilon_2^*, ..., \epsilon_r^*]$. Conventionally, whenever a device has a packet to transmit, it needs to locate and synchronize to a suitable BS based on its broadcast information. When the number of active devices is larger than $L_r^* N$, granting all devices access into the network will jeopardize the performance of all groups. In this case, the system may implement some form of access barring to limit the number of active devices in a given transmission frame. A simple example is given in the proposition below.

Proposition 4.4. *Let $\frac{N}{K}$ be the system load at the beginning of a transmission frame. These devices are blocked with a probability of b where b is calculated below as*

$$b = 1 - \min\left(1, L_r^*(\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\epsilon}^*)\frac{N}{K}\right). \tag{4.13}$$

Given $\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\epsilon}^*$ and N, the system can guarantee the required QoS requirements for at most $L_r^*(\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\epsilon}^*)N$ devices. Proposition 4.4 shows that when $K \leq L_r^*(\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\epsilon}^*)N$, all active devices are allowed access into the system ($b = 0$). On the other hand, when $K > L_r^*(\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\epsilon}^*)N$, an average of only $(1-b)K = L_r^*(\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\epsilon}^*)N$ active devices are allowed access into the system. Otherwise, the system will be overloaded and will fail to satisfy the QoS requirements. This access barring technique is often referred to as Dynamic Access Barring (DAB) as the probability is updated in each transmission frame based on the current load and is not fixed a priori. It is worthy of noting that ACB schemes [99] can also be implemented to block more devices of the less important applications rather than all applications equally.

4.5 Design of Reliable RA Schemes for a Finite Number of Devices

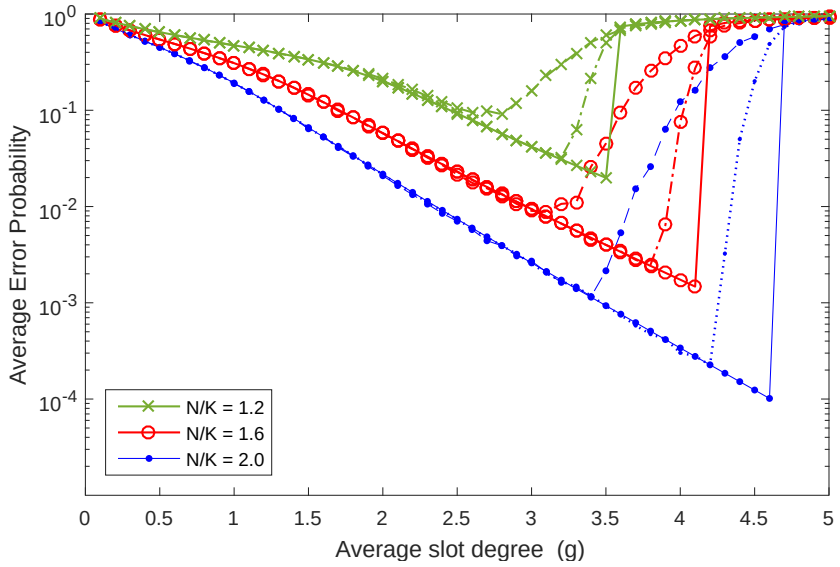

Figure 4.6 – Average error probabilities achievable by different access probabilities and different loads. The solid lines represent the error probabilities calculated from the AND-OR tree expressions. The dashed lines and the dash dotted lines represent the error probabilities obtained from simulations for $N = 200$ and $N = 2000$, respectively.

When concerned with maximizing the system reliability, i.e., minimizing the device resolution error probabilities, the inaccuracy of the asymptotic AND-OR tree results for finite values of K and N has been noted in previous works [46]. In this section, we elaborate on this discrepancy and propose a guideline that enables us to use the AND-OR tree expressions to design the access probabilities for a finite number of devices. We start off by considering only a single group ($r = 1$) of K devices transmitting with an access probability of $\frac{g}{K}$, where $g = g_1^{(1)} > 0$ is the average number of devices that access a given time slot.

Let us denote by $\epsilon\left(g, \frac{K}{N}\right)$ the average probability of device resolution error $\epsilon_1^{(1)}$ (ref. Section 4.2) when the system load is $\frac{K}{N}$ and the access probability is $\frac{g}{K}$. Here, we add the arguments g and $\frac{K}{N}$ to distinguish between the error probabilities achievable under

different system loads and using different access probabilities. It is also to emphasize that they are the only parameters necessary to calculate the average probability of device resolution error in the asymptotic case. We calculate $\epsilon\left(g, \frac{K}{N}\right)$ from Proposition 4.1 under different settings and plot the results in Figure 4.6 along with the simulation results for $N = 200$ and $N = 2000$. Let us denote by g^* the corresponding values of g at the extrema points, i.e., $g^* = \arg\min_g \epsilon\left(g, \frac{K}{N}\right)$. We notice that for both values of N there is a significant mismatch between the simulation results and the AND-OR expressions in the region $[g^* - \sigma, g^* + \sigma]$, where σ is a positive decreasing function of N, i.e., $\sigma \to 0$ when $N \to \infty$.

Moreover, we also observe that the error probabilities for all loads decrease gradually as g approaches the extrema points g^*. On the other hand, the error probability curves beyond that point become sharply increasing. For example, at $N/K = 1.2$, the error probability jumps from 0.02 to 0.61 when g increases from 3.49 to only 3.50. That is, in that region, the achievable error probability is very sensitive to the variations in g. Now, recall from Figure 4.1a, that the probability of an edge connecting a pair of VN and CN is Bernoulli distributed with a success probability of $\frac{g}{K}$. Therefore, the sum of edges in a bipartite graph is a Poisson distributed random variable with an average of gN. Accordingly, we can generate an infinite number of random bipartite graphs given the parameters K, N, and g. Let us denote by X the sum of edges of a random bipartite graph, and let $G = \frac{X}{N}$. Then, $\frac{G}{K}$ denotes the effective access probability of this bipartite graph. It is straightforward to see that G is also a Poisson distributed random variable with an average of g/N. Thus, the variations in G and the effective access probability of the graph increases with the decrease in N. For example, for $g \leq 4$, the standard deviation of G would be approximately 0.14 and 0.04 for $N = 200$ and $N = 2000$, respectively. Therefore, a system with a finite number of devices operating at the optimal points g^* will exhibit large variations in system performance, leading to the discrepancies observed between the actual average system performances and that predicted by the AND-OR tree in Figure 4.6.

Based on all the above, it is now clear why the optimal values of g^* for a finite

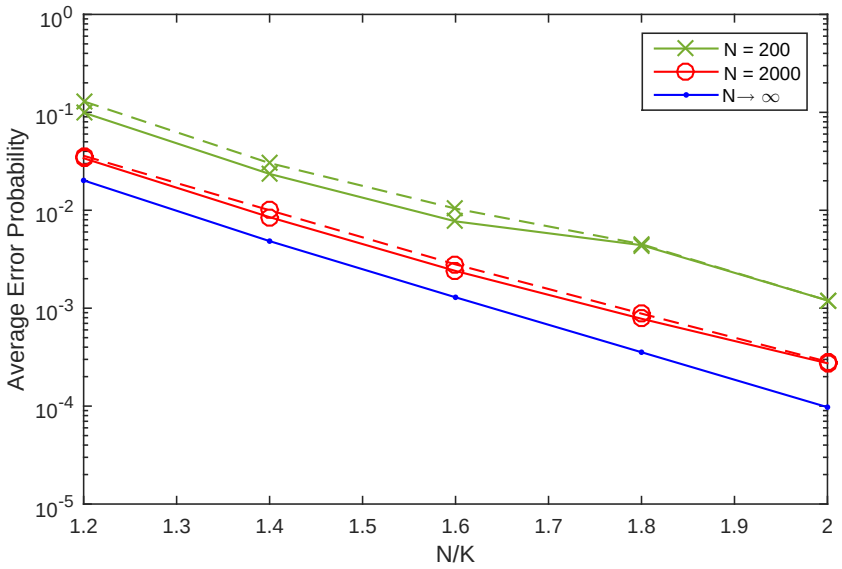

Figure 4.7 – Average error probabilities achievable by the proposed design guideline for $c = 10$. The solid lines represent the minimum error probabilities achievable in simulations and the dashed lines represent the error probabilities achieved by the design.

number of devices cannot be calculated directly from the AND-OR tree expressions. This behavior was noted in [46] and was avoided by conducting a simulation-based search of the optimal access probabilities. While this approach may be acceptable for a small number of devices, it becomes very tedious with the increase in the number of devices as well as the number of groups. We find that a much simpler way to find the optimal access probabilities for a given number of devices K and number of time slots N is searching for the value of g, which minimizes the average of all three values: $\epsilon\left(g-\sigma, \frac{K}{N}\right)$, $\epsilon\left(g, \frac{K}{N}\right)$ and $\epsilon\left(g+\sigma, \frac{K}{N}\right)$, where $\sigma = c\sqrt{\frac{g}{N}}$ for some positive constant c. Here, c determines the accuracy of the search. The accuracy of our proposed design is shown in Figure 4.7. We make note that was not necessary in the previous section as we were concerned with minimizing the number of transmissions rather than the error probabilities. In other words, we were not concerned with operating at the extrema points.

For the sake of completeness, we extend our design to the case of multiple QoS

requirements with 3 groups of devices. We set the acceptable average probability of device resolution error to 10^{-3} for all groups. Furthermore, for the sake of energy efficiency, we limit the average number of transmissions and only consider ranges of $0 \leq g_i^{(s)} \leq 4$. In Figure 4.8, we consider the case where all the groups in the system contain the same number of devices. We plot the achievable error probabilities with access probabilities designed directly from the AND-OR tree expressions and those designed using the aforementioned guideline. First of all, we observe the biggest mismatch between the two approaches for $\mathcal{C}_1$, whose minimum average probability of device resolution error is above the required threshold. This validates the importance of our proposed guideline in determining suitable points of operation.

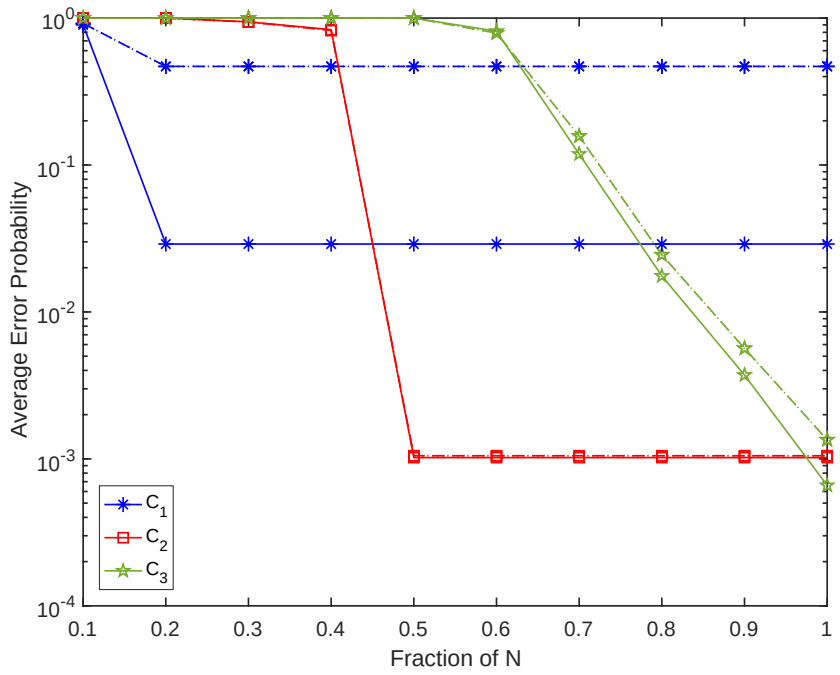

Figure 4.8 – Achievable probabilities of device resolution error by each scheme for $N/K = 2$, $\boldsymbol{\beta} = [0.2, 0.5, 1]$ and $\boldsymbol{\alpha} = [\frac{1}{3}, \frac{1}{3}, \frac{1}{3}]$. The solid lines and dotted lines correspond to the ACK-All scheme with c = 10 and c = 0, respectively. The dashed lines correspond to the ACK-Group scheme with c = 10.

Interestingly, we observe that both schemes demonstrate the same performance. This validates the results of Proposition 4.3 that non-separate transmissions are only pos-

sible when the load per each group is larger than the given bound. Otherwise, the optimal performance converges to that of the ACK-Group scheme. In Figure 4.8, we consider the same setup but for the case where the number of devices in each group is almost proportional to its delay requirement. In this case, the load per group is significantly larger than the bound in Equation 4.11, which allows for the sharing of resources. We plot the achievable performance of each group for different loads. For N/K approximately larger than 2, the ACK-All scheme can significantly improve the performance of $\mathcal{C}_3$ while guaranteeing $\mathcal{C}_1$ and $\mathcal{C}_2$ their required thresholds. Such an assumption on the sizes of the groups is practical when considering a Poisson packet arrival model where all devices arrive at the same rate. As devices with tighter delay requirements are served faster than others, the average number of queuing devices in $\mathcal{C}_i$ at the beginning of each transmission frame will be always less than those in $\mathcal{C}_j$, for $i < j$.

4.6 Performance Evaluation

4.6.1 LTE Setting

In this section, we evaluate the performance of RA in an LTE-based setting. In each transmission frame, a random number of resource blocks (RBs) is assumed to be available for M2M communications [54]. In LTE, a RB is the smallest radio resource unit that can be allocated to a device. It is made up of one time slot and one sub-channel. Thus, the incorporation of RA into an LTE-based setting requires an efficient resource management scheme of the two-dimensional resources. Although this is outside the scope of our work, we show that even with a simple setup, RA can still score better performance over coordinated access. The total number of RBs is assumed to be a uniformly distributed random variable with a predetermined mean. These RBs are divided between the RACH and the data channels. We assume that we can construct 8 preambles from each RB allocated to the RACH. In general, the number of preambles that can be constructed from one RB depends on the cell radius,

detection requirements and timing estimation accuracy [54]. We also assume that we can transmit one packet in each RB allocated to the data channels.

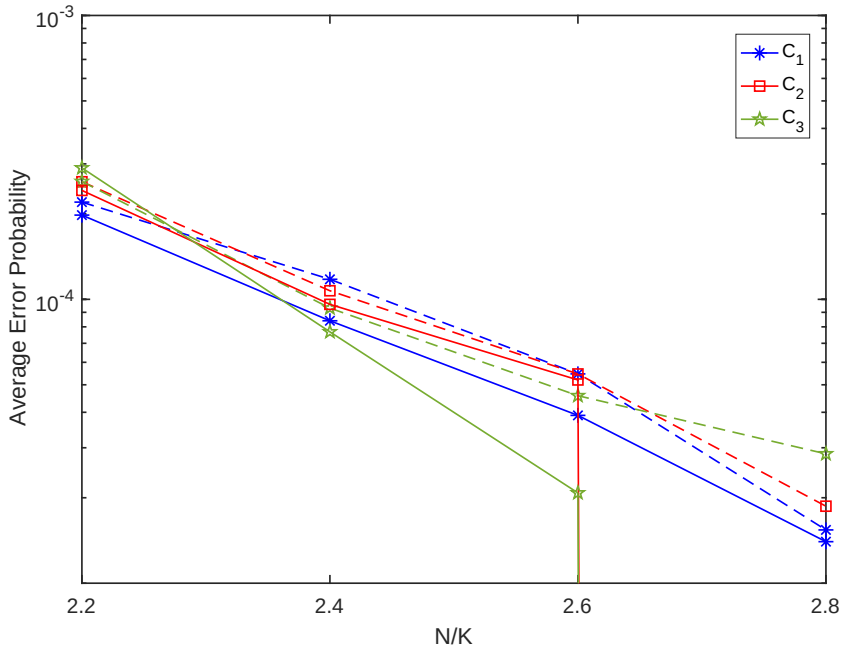

Figure 4.9 – Achievable probabilities of device resolution error by each scheme for $\boldsymbol{\alpha} = [0.2, 0.3, 0.5]$. The solid lines correspond to the ACK-All scheme, and the dashed lines correspond to the ACK-Group scheme with c = 10.

As access barring is said to be the best currently available solution to the RACH overload problem and has been used in the standardization process of LTE-A [117], we consider two schemes based on DAB as benchmarks. In the first scheme, we assume that the number of RBs allocated to the RACH in each transmission frame is fixed; thus, the number of available preambles is also fixed. However, for fair comparison, we still consider that the total number of RBs is a random variable; thus, the number of data channels in this case will vary from frame to frame. In the second scheme, we assume that the percentage of the RBs allocated for M2M communication in each transmission frame is fixed. Therefore, both the number of preambles and the number of data channels will vary from frame to frame. The blocking probability is dynamic in the sense that it changes from frame to frame based on the estimated load. The

blocking probability is calculated in the same way as in [54] to maximize throughput. Devices that are blocked or are unsuccessfully resolved are allowed to retransmit in the following frames. Nevertheless, as recent works [51] have shown that access barring and even DAB will not suffice as a stand-alone solution in future cellular networks, we consider the work of [54] which combines DAB with the dynamic allocation of RBs to the RACH and data channels as another benchmark. That is, the available RBs are allocated dynamically to each of the RACH and the data channels based on the estimated load to maximize throughput.

For RA, we use N to denote the number of RBs in a frame which is not necessarily the number of time slots. In fact, each transmission frame is assumed to be fixed in time duration, and the Poisson packet arrival rate λ is defined as the number of packets per frame. The BS is assumed to be capable of estimating the load in each frame. Therefore, each frame is the same as that in Section 4.2, with the total number of devices being the sum of new arrivals and unsuccessful transmissions from previous frames. Packets arriving during a frame are backlogged and wait for the next frame to start. DAB is also incorporated here to maximize the throughput in each frame. Figure 4.10 shows the throughput of the system defined as the average number of resolved devices per transmission frame. In each transmission frame, the number of RBs reserved for M2M communications is a uniformly distributed random variable varying from 0 to 100. Simulation results are plotted for different packet arrival rates. The average number of resolved devices is shown to increase for all schemes with the increase in λ until it reaches a certain saturation point. We observe that RA can achieve a higher throughput and, thus, can support a larger packet arrival rate.

When the throughput saturates, the system is said to be unable to service all active devices and the blocking probability will start to increase. As more and more devices are barred from access, the expected delays increase, and the system becomes unstable. We note that the system is said to be stable provided that the expected delay experienced by the devices is bounded below a certain threshold. This is shown in Figure 4.11, where the expected delay is expressed as the number of transmission frames. We can see that again RA can support a larger packet arrival rate while

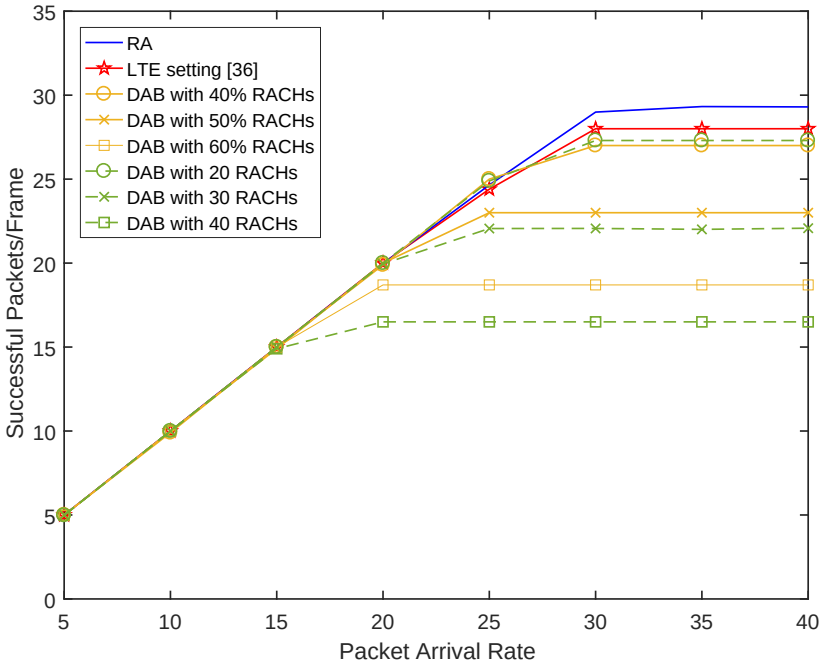

Figure 4.10 – Number of successfully recovered packets per frame for RA, DAB and the dynamic resource allocation scheme in [54]

guaranteeing smaller latencies. Finally, in Figure 4.12, we show the capacity of these schemes for different average number of RBs. The capacity is defined as the maximum throughput that can be stably supported. The capacity of RA is shown to be the largest when the number of RBs available is sufficiently large.

4.6.2 Practical Considerations

We now shed light on a number of important practical considerations that arise with the considered framework.

Load Estimation

We refer the readers to the work [118] for a load estimation algorithm for the case of batch arrivals. Authors in [118] claim that their algorithm can be readily extended

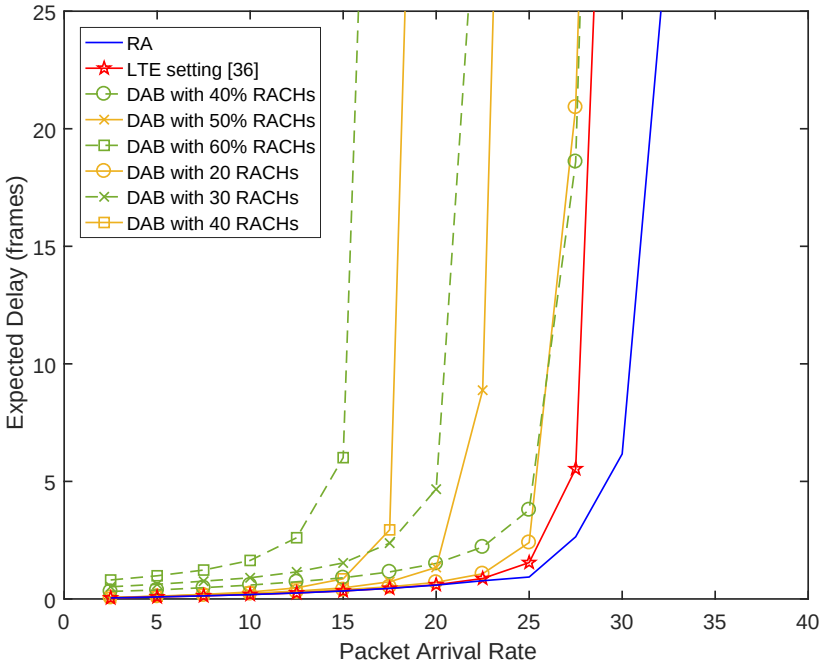

Figure 4.11 – Expected delay in number of frames for RA, DAB and the dynamic resource allocation scheme in [54]

to the case of Poisson packet arrivals, yet still admit that the evaluation complexity can be high. For that, we provide the readers here with a simpler load estimation algorithm that appears to work well for the selected case studies. In a nutshell, we can incorporate the number of resolved devices in each frame into the estimation. Let $K[i]$, $K_s[i]$ and $b[i]$ be the estimated number of participating devices, the number of successfully resolved devices and the access barring probability, in frame i, respectively. Then, the estimated number of participating devices can be expressed as $K[i] = \lambda + ((1 - b[i-1])K[i-1] - K_s[i-1]) + b[i-1]K[i-1]$, where the first, second and third terms correspond to the new packet arrivals in frame i, the unsuccessfully recovered packets in frame $i-1$ and the barred packets in frame $i-1$, respectively. In Figure 4.12, we show that the performance degradation due to inaccurate load estimation is quite negligible. This estimate can be further improved by considering the statistical information of the previous singleton, collision, and idle slots. Performance can also be improved by designing the system for a value of $K(1+\rho)$, where ρ is the

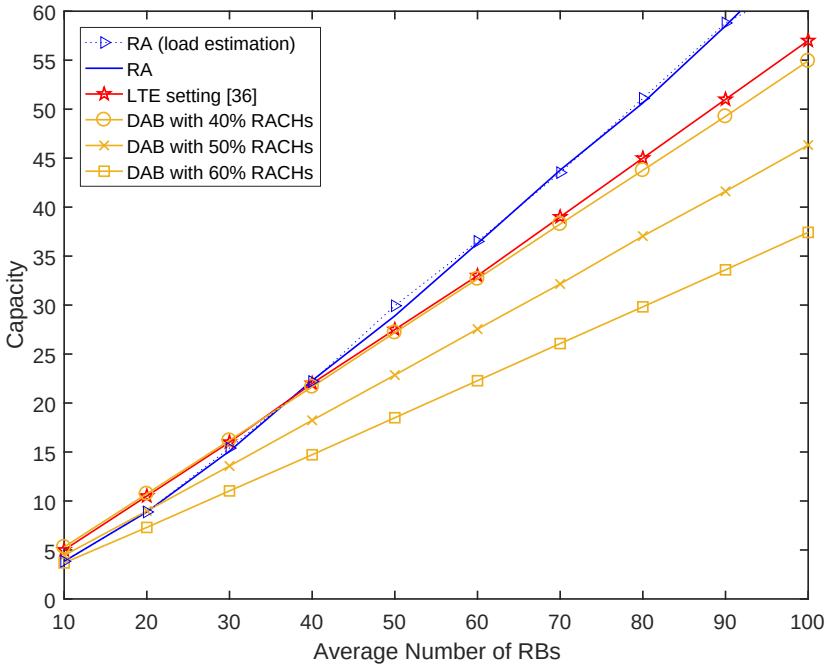

Figure 4.12 – Maximum achievable throughput for different average number of RBs.

fractional offset error in estimation.

Finally, we note that for all codes-on-graph [114, 115, 42] and random access schemes [41] with iterative decoding, there exists an upper bound on the rate/load below which we can have a zero or close to zero error probability. The threshold effect dictates that the error probability becomes more and more sensitive on the load as the load approaches the threshold. The threshold effect is important to note when considering load estimation errors. It is important to have estimation errors small enough to maintain sufficiently small error probabilities.

Lossy Feedback Channel

So far, we have assumed a perfect feedback. However, in some cases, acknowledgements to successfully resolved devices might be lost in the network. These unacknowledged devices will retransmit in the following sub-frames/frames assuming their packets have not been successfully recovered yet. This will induce a dynamic behavior

in the system and can affect the stability of the system. We investigate the impact of lossy feedback channels on the system performance in Figure 4.13. We can see that with an imperfect feedback channel, the losses in performance are minimal provided that the losses in acknowledgements are relatively small (< 0.01), as is generally assumed. Moreover, knowing the feedback channel state, one may redesign the system to a target error probability equal to the product of that of the SIC and that of the feedback channel.

We refer the readers to the work in [17] for a study on the reliability of control information and techniques to increase this reliability. In addition, we emphasize that some of the newly proposed paradigms are considering open-loop communications, where packets are not acknowledged. In this case, device transmissions are either limited in number or within a certain time frame, after which the devices either start transmitting another packet or become idle. Such settings have been shown to have significant gains in terms of network latency in comparison to closed-loop communications [119].

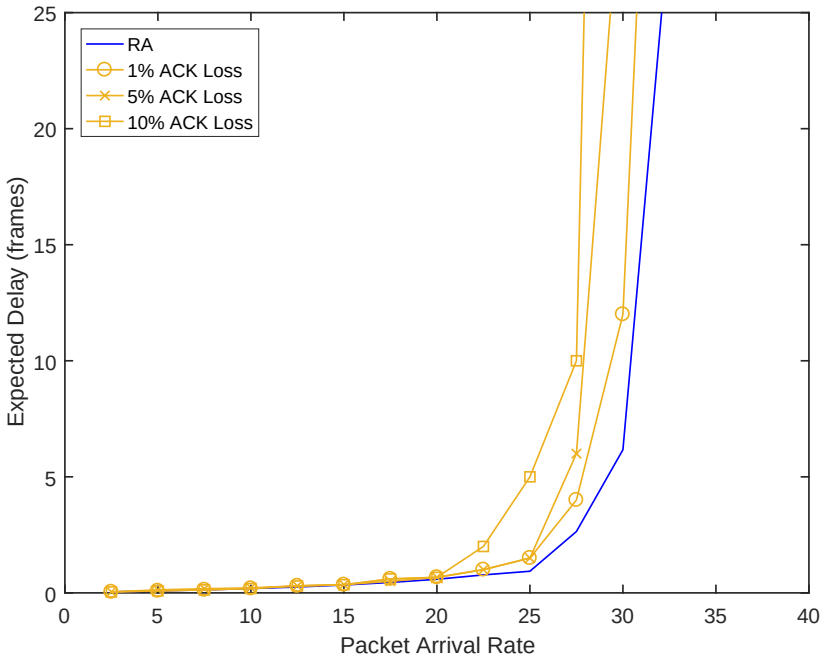

Figure 4.13 – The effect of lossy feedback channels on the stability of RA

Channel Estimation

Channel estimation is necessary in RA for two main reasons. First of all, based on the received power and known CSI, the BS can distinguish between idle slots, singleton slots and collision slots. Second of all, the received power needs to be large enough to allow correct decoding of the information for the given coding and modulation scheme at the receiver.
As we assume that packets can only be recovered from singleton slots, the detection of a singleton slot is based solely on the received power level. Therefore, issues with inter-symbol interference can be resolved using the same techniques used for point-to-point scenarios. We assume the inter-symbol interference caused by the overlapping between different packets in different slots will have little impact on the SIC performance due to the very low interference levels for small synchronization errors. To prevent such misalignment, some guard zone can be added at the beginning and the end of each symbol. Since we assume small synchronization errors, the duration of the guard zone in each symbol is also negligible. Furthermore, once a packet is recovered from a singleton slot, the BS can determine its timing offset. As the devices are assumed to be static for the duration of the transmission frame, the BS can use this information to accurately cancel the packet out from other collision slots.

Other RA schemes consider the capture effect where the BS can recover the packet from a collision slot provided that the SINR is larger than a certain threshold [48]-[C4]. For this setup, the system performance may be more sensitive to channel estimation errors. However, we do not consider the capture effect in this work. Finally, other practical considerations concerning the complexity and feasibility of the proposal in a real network are left for future work to be addressed in a more experimental approach.

4.7 Chapter Summary

In this chapter, we proposed a random access scheme with QoS guarantees for a heterogenous M2M communication network. We considered two transmission schemes:

ACK-All and ACK-Group. The ACK-All scheme allows for the simultaneous transmissions of different device groups over the same resources. Meanwhile, the ACK-Group scheme assumes that devices from different groups transmit over different resources. We drew an analogy between our proposed scheme and the codes on graph, and we derived the expressions for the average probability of device resolution for each of these two schemes based on the AND-OR tree. We showed the accuracy of these expressions in calculating the error probabilities and proposed a guideline to design the access probabilities in practical M2M settings. We showed that non-separate transmissions are only beneficial when the loads of the groups with tighter deadlines are relatively low to enable them to share their resources with the remaining groups. Otherwise, we showed through analysis and simulations that ACK-Group is optimal. Finally, we showed that the proposed scheme is superior to coordinated access schemes when the number of active devices and available resources is large enough.

Chapter 5

Grant Free Massive NOMA

5.1 Chapter Introduction

5.1.1 Chapter Background

One of the most common uncoordinated access schemes is the slotted ALOHA protocol of which many variants have been proposed over the past decade that aim at improving the system throughput as discussed in Section 2.2. However, as these protocols are orthogonal in nature, i.e., transmissions take place over a set of non-overlapping resource units, the number of devices that can be supported is dependent on the number of available resource units. Moreover, from an information-theoretic perspective, orthogonal multiple access has been shown to be strictly sub-optimal for short packet transmissions [32]. The gap between the achievable sum-rate of orthogonal multiple access and the maximum sum-rate increases as the system load increases [32].

On the other hand, authors in [120] showed that the maximum sum-rate is achievable through NOMA and joint decoding. This is because NOMA allows multiple users to share the different resources, e.g., time, frequency, and space, either through power domain multiplexing or code domain multiplexing [121, 122, 123]. Thus, unlike orthogonal multiple access, overloading is possible at the expense of increased processing

complexity at the receiver [123]. As mMTC communications are uplink-oriented, this complexity is at the AP and is, thus, acceptable. NOMA allows multiple users to share time and frequency resources in the same spatial layer via power domain or code domain multiplexing.

For uncoordinated NOMA, the problem is that the set of active users as well as their respective channel conditions are not known a priori at the AP. Different approaches to solve this problem have been proposed. In [124], a message passing algorithm was proposed to jointly detect user activity and their data. This problem was also solved in [125] using compressive sensing. In their technique, the estimated user set in each slot depends on prior information from previous transmissions. This is valid when there exist temporal correlations between user transmissions. In [74], another user detection technique was proposed using a set of orthogonal pilot sequences which are chosen uniformly at random by the set of active users. As long as every pilot sequence is chosen by only one user, users can be accurately detected at the receiver side and their channel state information can be accurately estimated. However, even when the number of pilot sequences is very large, the probability that two or more users choose the same pilot sequence is non-zero. In this case, a collision is said to have occurred as these devices cannot be distinguished by the AP. While authors in [75] suggest that devices adopt some back-off scheme to resolve this collision, this implies that any collision will lead to the loss of all data including those devices that did not collide. This is very wasteful of resources.

5.1.2 Contributions

Motivated by these findings, this chapter investigates the performance of an uncoordinated massive NOMA scheme where user detection and channel estimation is carried out via pilot sequences that are transmitted simultaneously with the device's data. In particular, we investigate the performance of massive NOMA with collisions, which is missing from previous works in this area. If the AP is not able to resolve the collisions, they are regarded as interference. In this case, other devices that did not collide in

the given slot can still be recovered. This allows for a more practical transmission scheme along with a more rounded assessment of its performance and suitability for mMTC. All in all, this chapter provides the following three major contributions.

Grant-Free Massive NOMA Scheme

We consider an uplink grant-free NOMA setting where devices jointly transmit a randomly chosen pilot sequence along with their data. For this setting, there is always a non-zero probability that two or more devices choose the same pilot sequence. The receiver is only able to estimate their aggregate power. However, it is unable to distinguish the devices from one and another, and a collision is said to have occurred. In this work, we propose to treat these codewords as interfering signals at the AP. We derive the distribution of the number of collided devices and show that the aggregate interference power can be well approximated by a PPP. Finally, we present the characteristic function of the aggregate interference power, which is an essential parameter in the performance analysis of this system.

Outage Probability of Massive NOMA

For the proposed framework, we first consider the case where all the devices transmit at the same fixed code-rate. We derive the expression of the outage probability for the case of joint decoding and successive interference cancellation. The evaluation of the exact expression is shown to be daunting especially for the case of joint decoding. To overcome this problem, we propose a simplified expression and demonstrate its accuracy through simulations. Our results show that the optimal length of the pilot sequences scales linearly with the packet arrival rate. Our results also show that SIC achieves a similar performance as SJD while reducing the decoding complexity.

Throughput of Grant-Free Massive NOMA

We then consider the case where the devices transmit using rateless codes. In this case, the rate is determined on the fly and varies from slot to slot based on the system

load, received powers and interferers. The receiver stops transmissions by broadcasting a beacon when the throughput is maximized. We derive the expression for the maximum throughput for the case of joint decoding and successive interference cancellation. The evaluation of the exact expression is shown to be very complicated. Based on this, we propose a simplified expression and demonstrate its accuracy through simulations. Our results show that the maximum throughput of SJD is almost double that of SIC. However, we explain that the maximum throughput under SIC is achievable in practice whereas the existence of code books that can achieve the maximum throughput in the case of SJD is questionable.

5.1.3 Chapter Outline

The rest of this chapter is organized as follows. In Section 5.2, we present the system model and transmission schemes. In Section 5.3, we derive the distribution of the received power, interference power, and number of interferers that will prove useful in our analysis. More importantly, we show that the received power is Pareto-distributed and that the interference can be well approximated by a PPP. In Section 5.4, we derive the outage probability and throughput of NOMA under SJD. In Section 5.5, we derive the outage probability and throughput expressions under SIC. Numerical results and practical considerations are presented in Section 5.6 and Section 5.7, respectively. Conclusions are drawn in Section 5.8.

The notations used in this chapter are summarized in Table 5.1 for quick reference. Also, in this chapter, we denote by $f_X(x)$, $F_X(x)$, $\psi_X(\omega)$ and $\mu_X(n)$ the PDF, the Cumulative Density Function (CDF), the Characteristic Function (CF), and the n^{th} moment of X, respectively. The cardinality of the set $\mathcal{X}$ is denoted by $|\mathcal{X}| = X$. All logarithms are taken to the base 2, unless otherwise indicated. $C(x) = \log(1 + x)$ is the point-to-point Gaussian channel capacity with x denoting the SNR, $\Gamma(\cdot)$ is the Gamma function, and $j = \sqrt{-1}$. Finally, $\delta(x)$ is the indicator function such that $\delta(x) = 1$ if $x > 0$ and is zero otherwise.

Table 5.1 – Notation Summary

Notation	**Description**
M	Number of symbols in a time slot
K	Number of information bits in a packet
q	Number of symbols in a pilot sequence
$\mathcal{N}$	Set of transmitting devices
$\mathcal{L}$	Set of pilot sequences
$\mathcal{L}_s$	Set of singleton pilot sequences/layers/devices
$\mathcal{N}_\ell$	Set of devices that chose the ℓ^{th} pilot sequence
$\mathcal{Z}$	Set of devices in collision (interfering)
P_T	Maximum transmit power of a device
P_i	Received power of device i
$\hat{P}_i$	Received power of the device in ℓ^{th} singleton layer
ρ_ℓ	SINR of layer ℓ, $\ell \in \mathcal{L}_s$
$\hat{\rho}_\ell$	SINR of the ℓ^{th} singleton layer
R_c	Device code rate
R_f	Device effective rate

5.2 System Model

5.2.1 Overview

We model the location of the devices transmitting in any time slot as a homogeneous PPP on an annular region with minimum and maximum radii $d_{\min}$ and $d_{\max}$ [126, 127, 128, 129, 130], respectively. The AP is located at the center, and the average number of transmitting devices surrounding the AP per time slot is denoted by λ. Devices are considered to be static, and the channel is modelled as a block fading channel. That is, the devices' channel conditions remain constant for the duration of one packet and vary randomly and independently from one slot to the other. For reciprocal channels, each device can make use of the pilot signal sent periodically over the downlink channel by the AP to synchronize their timing to that of the AP. The impact of asynchrony is discussed in Section 5.7.

For each time slot, the AP initiates uplink transmissions through beaconing. To minimize signalling overhead, we assume that the beacon signal carries load information based on which the devices adjust their code rate. The received signal at the AP for a given time slot t can be expressed as:

$$\mathbf{y}^{(t)} = \sum_{i \in \mathcal{N}^{(t)}} \mathbf{x}_i^{(t)} + \mathbf{w}^{(t)}, \tag{5.1}$$

where $\mathbf{x}_i^{(t)}$ is the codeword transmitted by device i from the total set of transmitting devices $\mathcal{N}^{(t)}$ with a received power of $P_i^{(t)}$, and $\mathbf{w}^{(t)}$ is a circular symmetric white Gaussian noise with unity variance (could include inter-cell interference). We consider Rayleigh fading and a path-loss channel model. Thus, $P_i^{(t)} = |h_i^{(t)}|^2 d_i^{-\alpha} P_T$, where h_i, d_i, α and P_T denote the small scale fading gain of the i^{th} device, the distance between the i^{th} device and the AP, the path-loss exponent and the transmit power, respectively. In what follows, we consider tight latency requirements where no retransmission opportunities are available, i.e., if the transmission of the packet is not successful in one slot, the packet is dropped. Thus, for the remaining part of the chapter, we focus our analysis on a single time-slot and, thus, we drop the

superscripts.

In this work, each codeword is concatenated with a pilot sequence of length q symbols for user detection and decoding. Thus, each codeword is of length $M - q$ symbols. Assuming each device has a set of K information bits to transmit, the code rate per device is defined as

$$R_c := \frac{K}{M-q} \quad \text{bits/symbol.} \tag{5.2}$$

On the other hand, the effective rate per device is defined as

$$R_f := \frac{K}{M} \quad \text{bits/symbol.} \tag{5.3}$$

R_f represents the ratio of the number of information bits to the total number of symbols transmitted. Thus, the effective rate takes into consideration the redundancy incurred by the pilot sequence.

5.2.2 Transmission Scheme

In each slot, each device chooses a pilot sequence of length q symbols independently and uniformly at random from a set $\mathcal{L} = \{1, 2, ..., L\}$. We categorize the pilots sequence into three types. An idle pilot sequence is a pilot sequence which has not been chosen by any device. A singleton pilot is a pilot sequence chosen by only one device, and a collision pilot sequence is a pilot sequence chosen by two or more devices. Each pilot sequence ℓ is made unique to a specific code book $\mathcal{C}_\ell$ and, thus, acts as the device's signature. A device i that has chosen the pilot sequence ℓ encodes its data $\mathbf{b}_i$ into a codeword $\mathbf{x}_i = f_\ell(\mathbf{b}_i)$.

In practice, the set of received pilot sequences is determined by the AP via the power delay profile which is constructed by performing cross-correlations between the known pilot sequences and the received signals and averaging the absolute square values of the created channel impulse responses [131]. This is based on the fact that the auto-correlation of orthogonal pilot sequences can be approximated by a delta function, and its cross-correlation with other pilot sequences yields all-zero sequences [131].

Thus, it is essential that the chosen set of pilot sequences have good auto and cross-correlation properties, e.g., Zadoff-chu [132], Golden codes [133], m-sequences [134]. In what follows, we assume perfect pilot sequence detection and power estimation, i.e., perfect device detection and channel estimation.

Some examples are illustrated in Figure 5.1. In Figure 5.1a, the four transmitting devices choose unique pilot sequences (1, 2, 4 and 8). Thus, there are no collisions. The received powers are estimated as P_1, P_2, P_4, and P_8. The remaining pilot sequences (3, 5, 6, 7, 9 and 10) are idle pilot sequences. In Figure 5.1b, three devices choose the first pilot sequences. Devices that choose the same pilot sequence also choose the same code book. These devices cannot be distinguished by the AP and are said to have collided. We will refer to each code book as a layer. Thus, these devices have collided over the first layer. In what follows, we assume that the AP can distinguish between collision layers and singleton layers. However, for collision layers, the AP does not know the number of colliding devices and is, thus, unable to decode the collision layers. Instead, the AP treats them as interference whose aggregate power can be estimated from the power delay profile. The feasibility of these assumptions will be discussed in Section 5.7.

We denote by $\mathcal{N}_\ell$ the set of devices that have chosen the ℓ^{th} pilot sequence of size N_ℓ. For a sufficiently large number of pilot sequences, the random variables $N_1, N_2, ..., N_L$ are Poisson random variables with average λ/L. This is a practical assumption as the length of the pilot sequences need to be large enough for the AP to accurately calculate the received powers, and the number of pilot sequences increases with its length [132]. The set of devices that have collided is denoted by $\mathcal{Z} = \{i \Big| i \in \mathcal{N}_\ell, N_\ell > 1, \forall \ell \in \mathcal{L}\}$. Moreover, the set of singleton layers is denoted by $\mathcal{L}_s = \{\ell \Big| N_\ell = 1, \forall \ell \in \mathcal{L}\}$. Then, the equivalent received signal as seen by the AP can be expressed as

$$\mathbf{y} = \sum_{n \in \mathcal{N}_\ell, \ell \in \mathcal{L}_s} \mathbf{x}_n + \sum_{n' \in \mathcal{Z}} \mathbf{x}_{n'} + \mathbf{w}, \tag{5.4}$$

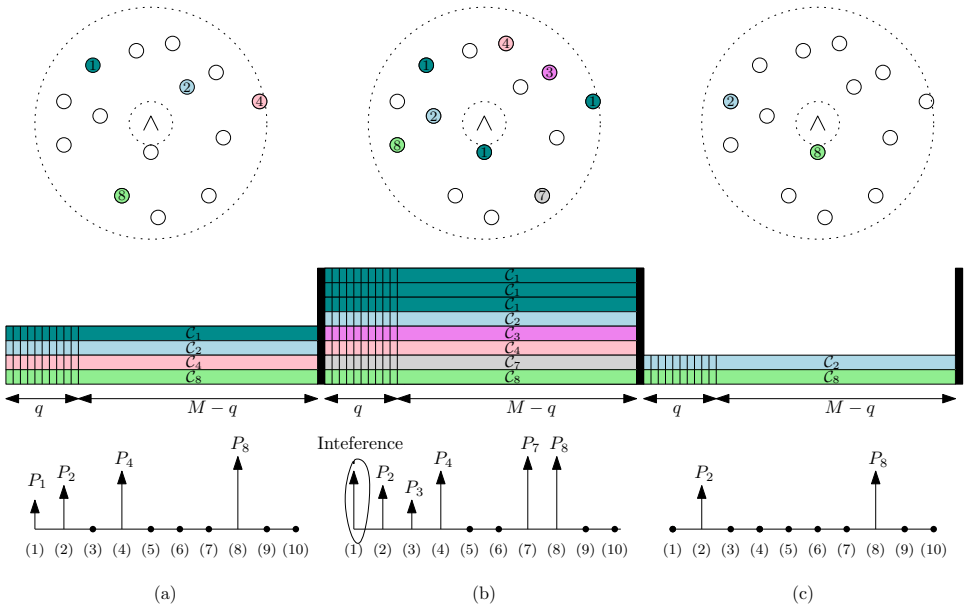

Figure 5.1 – Different case scenarios of a network with $L = 10$ pilot sequences: (a) $N = 4$, (b) $N = 8$ and (c) $N = 2$.

where $\mathbf{x}_i = f_\ell(\mathbf{b}_i)$ for $i \in \mathcal{N}_\ell$. The received SINR of the singleton layers is given as

$$\rho_\ell = \frac{P_i}{I + \sigma^2}, \quad \text{for } i \in \mathcal{N}_\ell, \ell \in \mathcal{L}_s, \tag{5.5}$$

where I denotes the aggregate interference power caused by these collisions given as:

$$I = \sum_{n \in \mathcal{Z}} |h_n|^2 d_n^{-\alpha} P_T. \tag{5.6}$$

It is worthy of pointing out the main differences between this transmission scheme and that of LTE. In LTE [54], the set of orthogonal sequences used for user detection are called preambles. Users transmit their chosen preambles separately from their data on a dedicated random access channel, and the AP allocates resources on the uplink data channel for each of the detected preambles. A collision occurs when two or more devices choose the same preamble. These devices will transmit their packets over the same time-frequency resources, and the AP will not be able to decode them correctly. Although many techniques have been proposed to resolve these collisions, e.g. [63], LTE remains a form of coordinated orthogonal transmission scheme and,

thus, is not suitable for massive access as explained in Section I. More importantly, the control signals and the data are transmitted over separate channels at different times which was shown to be strictly sub-optimal for small payloads [17].

5.3 Preliminaries

In this section, we first derive the distribution for the received power of a single user randomly located in the cell. Then, for Poisson packet arrivals, we show that the colliding/interfering devices can be well approximated by a PPP. Based on this, we find the CF for the aggregate power I defined as $\psi_I(j\omega) := \mathbb{E}\left[e^{-j\omega I}\right]$. These parameters will be useful in evaluating the outage probability and the throughput of the considered uncoordinated NOMA setting.

5.3.1 Distribution of Received Power

For the case where devices always transmit with maximum transmit power P_T over a channel subject to Rayleigh fading and path-loss, the distribution of the received power of a singleton layer is given in the following corollary. We refer the readers to Section C.1 for the proof. We make note that the extension of this to Nakagami fading or any other fading distribution is straightforward.

Corollary 1. *For a device uniformly distributed in an annulus with a minimum radius of $d_{\min}$ and a maximum radius of $d_{\max}$, the CDF of the received power $P = |h|^2 d^{-\alpha}$ ($P_T = 1$) at the origin under Rayleigh fading is*

$$F_P(p) = 1 - \frac{p^{-\frac{2}{\alpha}}\Gamma\left[\frac{2}{\alpha}+1\right]}{d_{\max}^2 - d_{\min}^2} + \frac{d_{\min}^2}{d_{\max}^2 - d_{\min}^2}. \tag{5.7}$$

5.3.2 Aggregate Interference Power

We now proceed to find the distribution of the aggregate interference power I. For that, we first find the exact distribution of the number of interferers in the following

lemma, i.e., number of devices in collision Z. We refer the readers to Section C.2 for the proof.

Lemma 4. *Consider a total of L layers where packet arrivals over each layer are Poisson distributed with an average of λ/L. Given that the number of singleton layers is $L_s < L$, the probability of having a total of n packets collide in a given time slot is expressed in Equation 5.8.*

$$Pr\Big(Z=n\Big|L_s\Big) = \frac{\left(\frac{(L-L_s)\lambda}{L}\right)^n e^{-\frac{(L-L_s)\lambda}{L}}}{n!(1-\lambda e^{-\lambda})^{L-L_s}}\left(1+\sum_{c=1}^{\min\{L-L_s-1,n\}}\binom{L-L_s}{c}\left(-\lambda e^{-\lambda}\right)^c\frac{\left((L-L_s-c)\frac{\lambda}{L}\right)^{n-c}e^{-(L-L_s-c)\frac{\lambda}{L}}}{(n-c)!}\right). \tag{5.8}$$

Although Z devices are only Poisson distributed in space and not in number, our results in Figure 5.2 show that their aggregate power can be well approximated by a PPP. Accordingly, the distribution of the aggregate interference power I, given L_s, can be well approximated by the skewed truncated stable distribution [135], and its CF can be expressed as

$$\psi_I(j\omega) = e^{\gamma_I\Gamma(-a_I)\left((g_I-j\omega)^{\alpha_I}-g_I^{\alpha_I}\right)}, \tag{5.9}$$

where the parameters α_I, g_I and γ_I determine the shape of the distribution. In particular, the parameters α_I and γ_I are related to the dispersion and the characteristic exponent of the stable distribution, respectively, and the parameter g_I is the argument of the exponential function used to smooth the tail of the stable distribution. These parameters are found through the method of the cumulants [135] and are given as

$$\alpha_I = \frac{2}{\alpha}, \tag{5.10}$$

$$g_I = \frac{\kappa_I(1)(1-\alpha_I)}{\kappa_I(2)}, \tag{5.11}$$

$$\gamma_I = \frac{-\kappa_I(1)}{\Gamma[-\alpha_I]\alpha_I \left(\frac{\kappa_I(1)(1-\alpha_I)}{\kappa_I(2)}\right)^{\alpha_I - 1}}. \tag{5.12}$$

Here, $\kappa_I(n)$ denotes the n^{th} cumulant of the interference power and is given as

$$\kappa_I(n) = \frac{2\lambda_Z}{n\alpha - 2} \frac{d_{\min}^{2-n\alpha} - d_{\max}^{2-n\alpha}}{d_{\max}^2 - d_{\min}^2} \mu_v(n) P_T^n, \tag{5.13}$$

where $v = |h|^2$ is a chi-squared distributed random variable under Rayleigh-fading, $\mu_v(n)$ is its n^{th} moment, and $\lambda_Z(L_s) := \mathbb{E}[Z \Big| L_s]$.

Finally, the inversion theorem [135] dictates that the CDF of I can be computed as

$$F_I(x) = \frac{1}{2} - \frac{1}{2\pi} \int_0^\infty \mathrm{Re}\left\{ \frac{\psi_I(-j\omega)e^{j\omega x} - \psi_I(j\omega)e^{-j\omega x}}{j\omega} \right\} \mathrm{d}\omega, \tag{5.14}$$

and the PDF of I can be computed as

$$f_I(x) = \frac{1}{2\pi} \int_0^\infty e^{-j\omega x} \psi_I(-j\omega) d\omega. \tag{5.15}$$

5.4 Performance of Massive NOMA with Successive Joint Decoding

In this section, we consider the case where all the singleton layers are jointly decoded at the receiver using a maximum likelihood decoder. Although jointly decoding all of the singleton layers is optimal in terms of the common outage event, i.e., the probability that at least one layer decoded incorrectly, it is not optimal in terms of the individual outage event, i.e., the probability that one layer is decoded incorrectly. Therefore, we consider the successive joint decoder (SJD) where a subset of the layers can be decoded jointly while regarding the remaining layers as interference. For that, we derive the expressions for the outage probability and the throughput. For ease of notation, we define $\hat{\mathbf{P}} := [\hat{P}_\ell]_{1 \le \ell \le L_s}$, where $\hat{P}_\ell$ is the received power of the device in

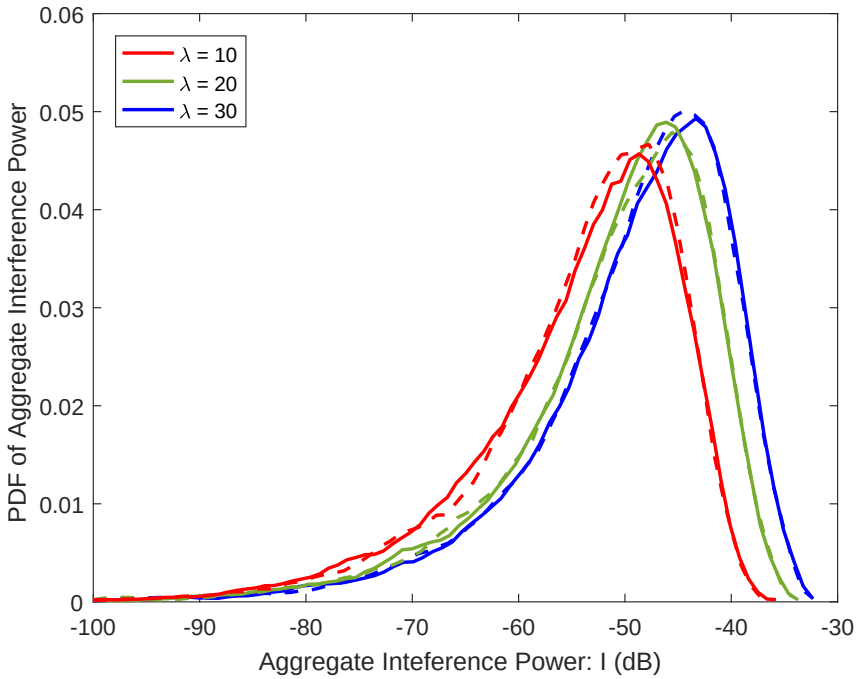

Figure 5.2 – Aggregate interference power for $L - L_s = 200$. The solid lines correspond to the actual distribution, and the dashed lines correspond to the Poisson distribution. Readers are referred to Table 5.2 for the remaining system parameters.

the ℓ^{th} singleton layer. Similarly, we define $\hat{\boldsymbol{\rho}} := [\hat{\rho}_\ell]_{1 \leq \ell \leq L_s}$, where $\hat{\rho}_\ell$ is the SINR of the ℓ^{th} singleton layer.

5.4.1 Outage Probability with SJD

Consider the case where all the devices encode their data using a fixed rate code with code rate $R_c := \frac{K}{M-q}$. In this case, we are interested in characterizing and evaluating the outage probability of the system. That is, the probability that a device is not successfully decoded in a given time slot. Given L_s singleton layers with a received power vector $\hat{\mathbf{P}}$, subject to an interference power I, a successive joint decoder first tries to decode all L_s layers jointly. However, if the equal-rate point $(R, ..., R) \in \mathbb{R}_+^{L_s}$ is outside the L_s-dimensional capacity region defined by the received SINRs $\hat{\boldsymbol{\rho}}$, the AP will try to decode the strongest $L_s - 1$ devices by treating the weakest layer as

interference. The weakest layer is defined as the singleton layer with the smallest received power. Similarly, if the equal-rate point $(R, ..., R) \in \mathbb{R}_{+}^{L_s-1}$ is outside the $L_s - 1$-dimensional capacity region, the AP will try to decode the strongest $L_s - 2$ layers by treating the two weakest layers as interference. The process repeats until decoding is successful or until there are no more layers to decode. Now, consider the descending ordered set $\hat{P}_{(1)} \geq \hat{P}_{(2)} \geq \ldots \geq \hat{P}_{(L_s)}$.

For that, the outage probability is defined as the average number of singleton layers that cannot be decoded successfully. Given L_s singleton layers with a received power vector $\mathbf{P}$, subject to an interference power I, the fraction of layers in outage can be expressed as

$$\begin{aligned} &\epsilon_{\text{SJD}}(L_s, I, \hat{\mathbf{P}}) = 1 - \max_{0 \leq \ell \leq L_s} \frac{\ell}{L_s}, \\ &\text{s.t. } R_c \leq \min_{1 \leq i \leq \ell} \frac{1}{i} \log \left(1 + \frac{\sum_{c=\ell-i+1}^{\ell} \hat{P}_{(c)}}{\sum_{c'=\ell+1}^{L_s} \hat{P}_{(c')} + I + \sigma^2} \right). \end{aligned} \tag{5.16}$$

Here, ℓ denotes the maximum number of layers that can be decoded out of the L_s singleton layers. In general, the outage probability can be expressed as

$$\epsilon_{\text{SJD}} = 1 - \mathbb{E}_{L_s, I, \hat{\mathbf{P}}} \left[\sum_{\ell=0}^{L_s} \frac{\ell}{L_s} \Pr \left(\bar{\phi}_{L_s, L_s}, \ldots, \bar{\phi}_{\ell+1, L_s}, \phi_{\ell, L_s} \right) \right], \tag{5.17}$$

where $\phi_{\ell, L_s} =$

$$\delta \left(R_c \leq \min_{1 \leq i \leq \ell} \frac{1}{i} \log \left(1 + \frac{\sum_{c=\ell-i+1}^{\ell} \hat{P}_{(c)}}{\sum_{c'=\ell+1}^{L_s} \hat{P}_{(c')} + I + \sigma^2} \right) \right),$$

Thus, to find the outage probability, not only do we need to average over the different number of singleton layers as well as the different aggregate interference powers, we also need to average over the numerous realizations of the received power vector. Thus, computing the outage probability for uncoordinated multiple access is even more daunting than that for coordinated multiple access [136]. In what follows, we make use of the massive aspect of mMTC to simplify this expression.

Consider a set of L i.i.d. random variables $[X_i]_{1 \leq i \leq L}$, where $X_1 \geq X_2 \geq ... \geq X_L$.

Then, from ordered statistics [137], we have

$$\lim_{L\to\infty} \Pr\left(|X_{\lfloor yL \rfloor} - F_X^{-1}(1-y)| > \nu\right) = 0, \tag{5.18}$$

for all $y \in [0,1]$ and $\nu > 0$. Based on this, for an asymptotically large number of singleton layers, the following equality holds for all $\ell_1, \ell_2 \in [1, L_s]$.

$$\begin{aligned}\lim_{L_s\to\infty} \sum_{i=\ell_1}^{\ell_2} \hat{P}_{(i)} =& L_s \int_{\ell_1/L_s}^{\ell_2/L_s} F_P^{-1}(1-y)\mathrm{d}y \\ =& L_s \left(g(\ell_2/L_s) - g(\ell_1/L_s)\right),\end{aligned} \tag{5.19}$$

where $g(y) := \int F_P^{-1}(1-y)\mathrm{d}y$. For Rayleigh fading and a path-loss channel model, $g(y)$ is evaluated by integrating the inverse of (5.7) and is expressed in Equation 5.20.

$$g(y) = \frac{2\left((d_{\max}^2 - d_{\min}^2)y - d_{\min}^2\right)}{(\alpha-2)(d_{\max}^2 - d_{\min}^2)\left(\Gamma\left[\frac{2}{\alpha}+1\right]\right)^{\frac{-\alpha}{2}}\left((d_{\max}^2 - d_{\min}^2)y - d_{\min}^2\right)^{\frac{\alpha}{2}}}. \tag{5.20}$$

For the special case of Rayleigh fading only, we have

$$g_{\text{Ray.}}(y) = (1-y)\log_e(1-y) - (1-y). \tag{5.21}$$

Using this property, the complexity of computing the outage probability of massive NOMA in Equation 5.17 is significantly reduced as $\mathbf{P}$ converges to a deterministic value as $L_s \to \infty$. The expression for that is derived in the following lemma. We refer the readers to Section C.3 for the proof. We make note that in practice, we find that this property gives sufficiently accurate results for $L_s > 10$.

Lemma 5. *Consider a time slot of length M symbol durations. For a code rate R_c, L pilot sequences of length q and a packet arrival rate of λ packets, the outage probability of the massive NOMA system with SJD can be expressed as*

$$\epsilon_{SJD} = 1 - \mathbb{E}_{L_s}\left[\int_0^1 \frac{v}{2\pi}\int_0^\infty e^{-j\omega I_{SJD}^*(v)}\psi_I(-j\omega)d\omega\right], \tag{5.22}$$

where $I^*_{SJD}(v)$ is the solution to the equation below

$$L_s R_c - \min_{0<u\leq 1} \frac{1}{uv} \log\left(1 + \frac{g(v) - g(v-uv)}{g(1) - g(v) + \frac{I+\sigma^2}{L_s}}\right) = 0. \tag{5.23}$$

Here, $I^*_{\text{SJD}}(v)$ denotes the largest interference power for which the largest fraction of layers that can be decoded successfully is v. The integral in Lemma 5 averages over the aggregate interference power, and the expectation averages over the number of singleton layers.

5.4.2 Maximum Throughput with SJD

We now consider the case where devices can transmit as many coded symbols as necessary. This is feasible when devices use rateless channel codes to encode their information [138]. With rateless codes, the length of the codeword is determined on the fly and is adaptive to the load, received powers and interference power. Thus, the duration of the slot (M) as well as the devices' code rate (R_c) also varies from one slot to the other. Transmissions are stopped by the AP when the throughput is maximized through beaconing. The throughput is defined as the ratio of the number of successfully decoded information bits to the length of the transmitted codeword ($M - q$). This is determined by the AP and can be made known to the devices through beacons. Given L_s singleton layers, a received power vector $\hat{\mathbf{P}}$, subject to an interference power I, the maximum instantaneous throughput is given as

$$\zeta_{\text{SJD}}(L_s, I, \hat{\mathbf{P}}) = \max_{1\leq \ell \leq L_s} \min_{1\leq i\leq \ell} \frac{\ell}{i} \log\left(1 + \frac{\sum_{c=\ell-i+1}^{\ell} \hat{P}_{(c)}}{\sum_{c'=\ell+1}^{L_s} \hat{P}_{(c')} + I + \sigma^2}\right).$$

As the throughput varies from slot to slot, the maximum average throughput should be averaged over all possible values of L_s, I and $\hat{\mathbf{P}}$. However, from Equation 5.18, the maximum instantaneous system throughput under massive access can be expressed

from [136] as

$$\zeta_{\text{SJD}}(L_s, I, \hat{\mathbf{P}}) = \max_{0\leq v\leq 1} \min_{0<u\leq 1} \frac{1}{u} \log\left(1 + \frac{g(v) - g(v-uv)}{g(1) - g(v) + \frac{I+\sigma^2}{L_s}}\right). \tag{5.24}$$

Then, the average maximum system throughput under massive access is derived in the following lemma. We refer the readers to Section C.4 for the proof.

Lemma 6. *For L pilot sequences of length q and a packet arrival rate of λ packets, the average maximum system throughput of massive NOMA with SJD can be expressed as in Equation 5.25.*

$$\zeta_{SJD} = \mathbb{E}_{L_s}\left[\max_{0\leq v\leq 1} \min_{0<u\leq 1} \frac{1}{u} \int_0^\infty \frac{\psi_I(-s)}{s} \left(e^{-s((g(1)-g(v))L_s+\sigma^2)} - e^{-s((g(1)-g(v-uv))L_s+\sigma^2}\right) ds\right]. \tag{5.25}$$

5.5 Performance of Massive NOMA with Successive Interference Cancellation

As joint decoding is often infeasible in practice due to complexity constraints, we consider SIC in this section. In each stage of SIC, the strongest layer is decoded by regarding remaining layers as interference. Thus, we only need a single-user decoder. If the layer is decoded successfully, the decoded layer is subtracted from the received signal. We assume perfect SIC. In the second stage, the second strongest layer is decoded by regarding the remaining layers as interference. SIC stops when a layer is decoded unsuccessfully or when there are no more layers to decode.

5.5.1 Outage Probability with SIC

Given that the strongest $\ell-1$ layers have been decoded successfully, the ℓ^{th} strongest layer is decoded successfully if the following condition is true.

$$\frac{\hat{P}_{(\ell)}}{\sum_{c=\ell+1}^{L_s}\hat{P}_{(c)}+I+\sigma^2} \geq 2^{R_c}-1. \tag{5.26}$$

Thus, given that the strongest $\ell-1$ layers have been decoded successfully, the probability that the ℓ^{th} strongest layer is decoded successfully is given as

$$\Pr\left(I \leq \frac{\hat{P}_{(\ell)}-\left(2^{R_c}-1\right)\left(\sum_{c=\ell+1}^{L_s}\hat{P}_{(c)}+\sigma^2\right)}{2^{R_c}-1}\right).$$

From Equation 5.18, given L_s and I, the SINR at each stage of the SIC is deterministic under massive access. Based on this, the outage probability under SIC is derived in the following proposition. The definition of the outage probability is as given in Section 5.4.1, and the proof of the proposition is similar to that of Lemma 5.

Proposition 5.1. *For a code rate R_c, L pilot sequences of length q and a packet arrival rate of λ packets, the outage probability of massive NOMA with SIC can be expressed as*

$$\epsilon_{SIC}(R) = 1-\mathbb{E}_{L_s}\left[\int_0^1 \frac{v}{2\pi}\int_0^\infty e^{-j\omega I^*_{SIC}(v)}\psi_I(-j\omega)d\omega\right], \tag{5.27}$$

where

$$I^*_{SIC}(v) := \frac{F_P^{-1}(v)-\left(2^{R_c}-1\right)\left((g(1)-g(v))L_s+\sigma^2\right)}{2^{R_c}-1}. \tag{5.28}$$

Here, $I^*_{\mathrm{SIC}}(v)$ denotes the largest interference power for which the largest fraction of layers that can be decoded successfully is v.

Table 5.2 – System Parameters

Parameter	**Value**
Cell Radius	500 m
Reference Distance	50 m
Bandwidth	15 kHz
Transmit Power	23 dBm
Thermal Noise Density	-174 dBm/Hz
Receiver Noise Figure	3 dB
Path loss exponent	3.5

5.5.2 Maximum Throughput with Successive Interference Cancellation

We now consider the same setup in Section 5.4.2. Given L_s singleton layers with a received power vector $\hat{\mathbf{P}}$, subject to an interference power I, the maximum instantaneous throughput is given as

$$\zeta_{\text{SIC}}\left(\mathbf{P}, L_s, I\right) = Ls \max_{0<v\leq 1} v \log\left(1 + \min_{0<v'\leq v} \frac{F_P^{-1}(1-v')}{(g(1)-g(v'))L_s + I + \sigma^2}\right). \tag{5.29}$$

The average maximum system throughput is given in the following proposition. The proof of this proposition is similar to that of Lemma 6.

Proposition 5.2. *For L pilot sequences of length q and with a packet arrival rate of λ packets, the average maximum system throughput of massive NOMA under SIC can be evaluated from Equation 5.30.*

$$\zeta_{SIC} = \mathbb{E}_{L_s}\left[\max_{0<v\leq 1} vL_s \min_{0<v'\leq v} \int_0^\infty \frac{\psi_I(-s)}{s} e^{-s((g(1)-g(v'))L_s+\sigma^2)}\left(1 - e^{-sF_P^{-1}(1-v')}\right) ds\right]. \tag{5.30}$$

5.6 Numerical Results

In our simulations, we consider single-tone transmissions with the system parameters listed in Table II [139, 140, 141]. The number of pilot sequences L is determined

such that $1 - e^{-\frac{\lambda}{L}} = \beta$, where β denotes the probability that a device suffers from collision. We also assume that $q = L$ [132, 133, 134]. This means that the overhead associated with channel estimation and user detection scales linearly with the packet arrival rate. In what follows, we evaluate the performance of massive NOMA for different collision probabilities.

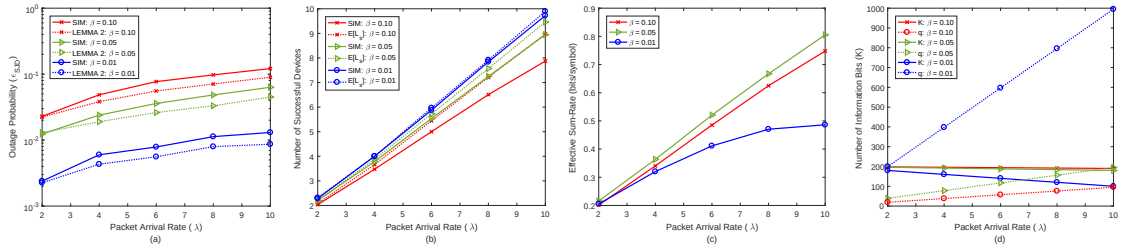

Figure 5.3 – Outage Probability for massive NOMA with SJD for different values of β. The code rate $\frac{K}{M-q}$ is equal to 0.1 and the slot duration is $M = 2000$ symbols.

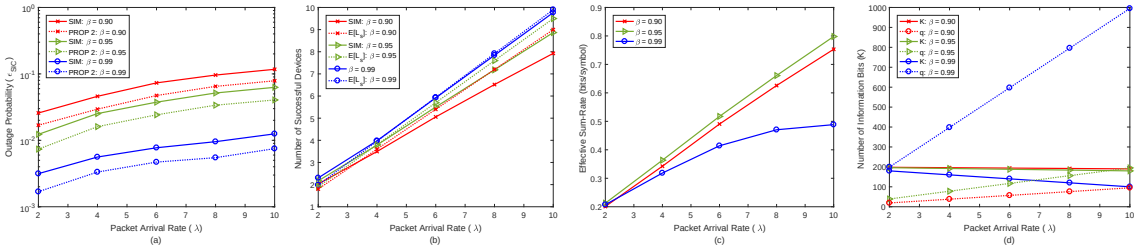

Figure 5.4 – Outage Probability for massive NOMA with SIC for different values of β. The code rate $\frac{K}{M-q}$ is equal to 0.1 and the slot duration is $M = 2000$ symbols.

First, we evaluate the outage probability for massive NOMA. Each time slot is of length $M = 2000$ [139] over which the devices transmit with a code rate of 0.1. Results for SJD and SIC are shown in Figure 5.3 and Figure 5.4, respectively. We first observe that the analytical results from Lemma 5 and Proposition 5.1(PROP 5.1) provide a good approximation of the actual performance obtained from Monte Carlo simulations. When comparing the performance for different collision probabilities, we find that the gap between the number of successfully decoded devices and the number of singleton layers becomes more significant as β increases for both SJD and SIC. This is because collisions in NOMA act as interference to the remaining singleton layers,

which leads to more outages.

In Figure 5.3c and Figure 5.4c, the effective sum-rate is evaluated as the product of the number of successfully decoded devices and the effective rate $\frac{K}{M}$. In general, if the number of pilot sequences is too small, the number of singleton layers will be low. Thus, the number of successfully decoded devices will be low as well. On the other hand, when the number/length of pilot sequences is too large, the overhead will be large. Thus, the effective rate will be low. Interestingly, it seems that the optimal value of L scales linearly with the packet arrival rate. Finally, in Figure 5.3d and Figure 5.4d, we plot the maximum payload sizes that can be transmitted and compare them to the overhead induced by the pilot sequences. In general, we can say that for applications with payloads of size 100-200 bits and a target outage probability of ≈ 0.01, we can simultaneously support up to 7 devices with SJD and up to 6 devices with SIC.

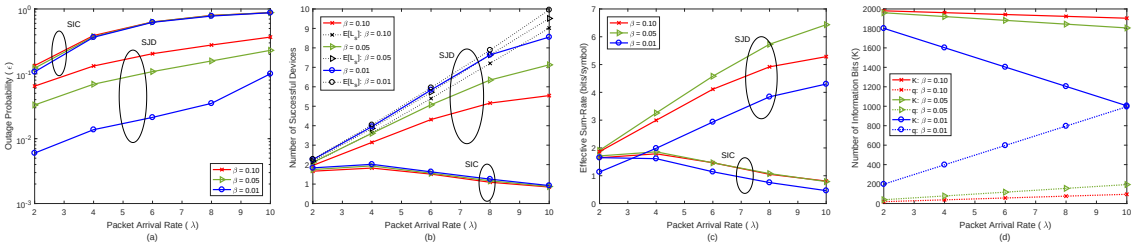

Figure 5.5 – Outage Probability for Grant-Free Massive NOMA with SJD and SIC for different values of β. The code rate $\frac{K}{M-q}$ is equal to 1 and the slot duration is $M = 2000$ symbols.

In Figure 5.5, we evaluate the outage probability for a larger code rate ($R_c = 1$). Here, we notice that the performance gap between SJD and SIC is more significant. In fact, from Figure 5.5b, we can see that SIC can barely decode any packet. However, although the outage probability of SJD is better, it is still too high for any practical use. Moreover, as mMTC are characterized by their low data rate transmissions, SIC seems to be a good candidate for our uncoordinated massive NOMA scheme when low processing complexity is needed.

Next, we evaluate the throughput for massive NOMA in Figure 5.6 and Figure 5.7.

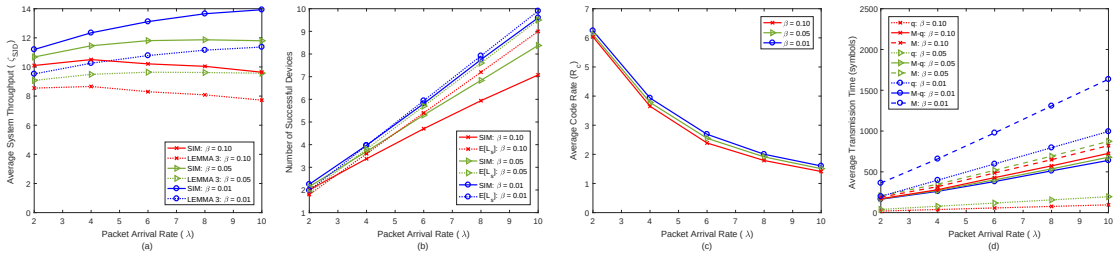

Figure 5.6 – Average system Throughput for massive NOMA with SJD for different values of β ($K = 1024$).

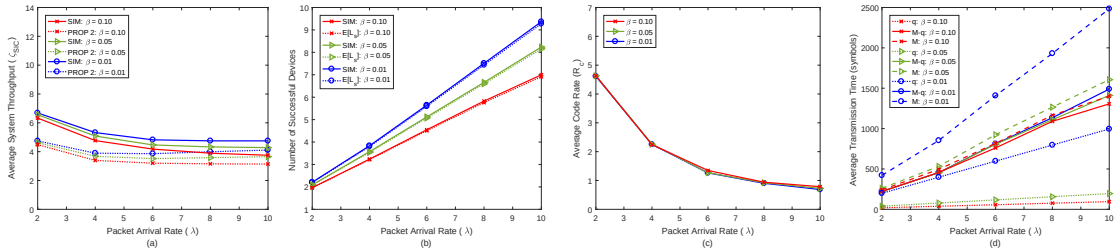

Figure 5.7 – Average system Throughput for massive NOMA with SIC for different values of β ($K = 1024$).

In this setup, each device has a set of 1024 information bits to transmit [139]. The number and length of pilot sequences is determined as before. We observe that the analytical results from Lemma 6 and Proposition 5.2 (PROP 5.2) also provide a good approximation of the actual performance obtained from Monte Carlo simulations. We also observe that the throughput gain of SJD is almost double of that of SIC, thus, SJD is in theory far more superior in this case. However, in the following section, we explain that the maximum throughput under SIC is achievable in practice whereas the achievability in the case of SJD is questionable. Finally, we make note that, in practice, M should not be larger than the coherence time. In this case, collision probabilities lower than 0.01 cannot be supported by this scheme as it will require longer pilot sequences.

5.7 Practical Considerations

5.7.1 Detection of Collision Layers and Achievability

For SIC, the detection of collision layers can be implemented through a simple algorithm. At first, the AP assumes that all the layers are singleton layers and orders the layers according to their respective received powers in descending order. In the first iteration, the AP attempts to decode the strongest layer. The AP can verify that the decoded information is correct through a simple CRC-check. If decoding is successful, the decoded signal is cancelled from the received signal. If not, the AP assumes that this is a collision layer. In the second iteration, the AP attempts to decode the second strongest layer while regarding the previously undecoded layer as interference. The AP can continue to follow the same steps until there are no more layers to decode. Detecting collision layers for SJD is also possible at the expense of a larger complexity, as the AP would need decode all possible subsets of layers while regarding the remaining layers as interference. When decoding finishes, the AP would choose the largest subset of successfully decoded layers.

It is straightforward to see that the derived bounds for SIC are achievable when capacity achieving point-to-point channel codes are used. However, for SJD, it is uncertain whether there exist channel codes suitable for grant-free coordinated access with joint decoding. That is mainly because these codes would not only have to adapt to the different channel conditions but also to the different channel loads. Finally, we would like to point out that there exist practical low-complex decoders such as belief-propagation-based decoders that can score performances close to that of maximum likelihood decoding.

5.7.2 Synchronization

In conventional communication systems, synchronization takes place beforehand. However, it has been well established that massive access control signals should be

minimal if not null. We would like to point out that, as most MTDs are stationary devices, their timing advance need not be communicated very frequently. Alternatively, time asynchrony can provide another contention unit by which devices can be distinguished from one another without having to increase the pilot sequence length and thus without introducing redundancy. Moreover, time asynchrony can help in detecting the number of devices in collision layers. This allows the opportunity to resolve collisions layers and decode the packets of the involved devices. However, the performance analysis for this case scenario is not straightforward as the maximum achievable rate with code-reuse is strictly lower than that with unique codes and is dependent on the codes used [142].

In general, the AP can feedback a timing advance signal in the acknowledgement. For that, the operator can either choose to operate either in a fully synchronous mode or allow the devices to introduce random asynchrony into their transmissions to create more contention units [143].

5.8 Chapter Summary

In this chapter, we considered a massive uncoordinated NOMA setting where devices choose pilot sequences from a predetermined set as their signature. Then, each device encodes its data using the pilot as the signature and transmits its selected pilot and data simultaneously with the rest of the devices. In our proposed scheme, a collision occurs when more than one device choose the same pilot sequence. The set of collided packets are treated as interference. For that, we show that the aggregate interference power can be well approximated by a PPP. Based on this, we derive the outage probability and the maximum system throughput under successive joint decoding (SJD) as well as successive interference cancellation (SIC) for a Rayleigh fading and path loss channel model. We verified the accuracy of our derived expressions via simulations. Our results show that SIC performs close to that of SJD in terms of outage probability for packet arrival rates up to 10 packets per slot. In terms of throughput, although SJD scores almost double the throughput gain of that of SIC,

we explained that this throughput might not be achievable with practical modulation and channel coding schemes.

Chapter 6

Conclusion

In this thesis, we studied the uncoordinated access of the emerging mMTC. We proposed several schemes with demonstrated improved throughput, energy efficiency, reliability as well as diverse QoS guarantees. This chapter provides a summary of the thesis content, a summary of the contributions and results, and sheds light on potential future directions.

6.1 Summary of Content and Results

In **Chapter 2**, we provided the background information of the main topics studied in this thesis. We first presented the fundamentals of the multiple access channel in terms of its definition, main components, and capacity limits. We also introduced the basic multiple access schemes used in previous and current wireless communications. Based on that, we distinguished between two main forms of access, namely, random access and coordinated access. The latter access scheme has been used in previous generations of cellular network. On the other hand, random access has been confined mostly to control channels so far. Then, we show that the characteristics of mMTC are significantly different than conventional human type communications for which cellular networks have targeted so far. In particular, the sporadic transmissions of short blocks of data by an enormous amount of power-limited MTC devices cannot

be supported efficiently and reliably over current cellular access networks. For that, there is a large consensus in this field that random access is more suitable. Finally, as IoT is forecasted to occupy a significant portion of future wireless communications, our study is focused on the analysis, design and optimization of random access for the heterogeneous applications of mMTC.

In **Chapter 3**, we proposed an SINR-based random access scheme where collisions are defined as the event of the received SINR of a packet dropping below the desired threshold. A packet in collision cannot be recovered directly unless the interference it suffers from is made known. This is possible through SIC. We drew an analogy between the packet recovery process in SINR-based RA using SIC and the iterative decoding process of codes-on-graph over the BEC channel. Based on this analogy, we proposed a tree-based analytical framework to track the evolution of the error probabilities in each iteration of the SIC process. We also derived the necessary conditions for convergence, i.e., the necessary conditions to achieve error probabilities close to zero after a large number of iterations. Based on the derived expressions, we solve for the optimal transmission probabilities via evolution-based strategies. The SINR-based model scores significantly larger throughput and can support loads larger than one. Finally, we extend this scheme to a CRN where the transmission probabilities of the SUs are optimized. Namely, their throughput is maximized while maintaining the interference to the primary network below an acceptable threshold. Our numerical results showed that our design is more reliable and more energy efficient than conventional RA schemes.

In **Chapter 4**, we proposed a multi-stage random access scheme that can guarantee the diverse latency requirements of the heterogeneous applications of mMTC. We proposed two different schemes where the devices are grouped based on their QoS requirements. In the ACK-All scheme, all groups transmit simultaneously over the same radio resources in all stages of the transmission frame. In the ACK-Group scheme, each group transmits in a distinct stage. We used the AND-OR tree to analyze the system performance characterized by the average probability of device resolution error for both schemes. Then, we used the derived expressions to design

schemes that can guarantee the QoS requirements of different groups with significantly high energy efficiency and high reliability. We showed that the proposed RA schemes can service a larger system load in comparison to coordinated access when the number of time-frequency resources is sufficiently large. Finally, we proposed a guideline that uses the derived expressions, suitable for asymptotically large networks, to optimize the transmission parameters for small networks. We demonstrated the simplicity and accuracy of our guideline in finding the optimal parameters in several case scenarios that would otherwise require an exhaustive search.

In **Chapter 5**, we proposed a massive uncoordinated NOMA scheme where user detection and channel estimation are carried out via orthogonal pilot sequences that are transmitted simultaneously with the devices' data. A collision event is defined as the event of two or more devices choosing the same pilot sequence. The receiver cannot directly detect collisions. It can only estimate the aggregate power of each detected pilot sequence. We propose practical schemes to implicitly detect these collisions to recover the data of the devices that have not collided. By treating these collisions as interference, we derive the maximum outage probabilities and throughput of uncoordinated NOMA under SJD and SIC. By demonstrating that the aggregate interference power can be well approximated by a PPP, and by assuming a massive access setting, we show how, the otherwise daunting, expressions can be evaluated in a simplified manner. We compare the two decoding schemes for different case scenarios. Our results show that SIC performs close to that of SJD in terms of outage probability for packet arrival rates up to 10 packets per slot. In terms of throughput, although SJD scores almost double the throughput gain of that of SIC, we explained that this throughput might not be achievable with practical modulation and channel coding schemes.

6.2 Future Work

To conclude this thesis, we list a number of promising research directions that follow from the work conducted herein.

6.2.1 Grant-Free NOMA with Short Packet Transmissions

We showed in Chapter 2 that RA is more suitable than coordinated access for mMTC as it incurs little or no overhead when the number of users is large and the transmitted packets are small. Furthermore, we showed that NOMA is capacity achieving whereas OMA schemes are strictly sub-optimal in this sense. Motivated by this, our future research work will focus on designing and optimizing practical uncoordinated NOMA schemes that can achieve, or at least approach, the capacity limits dictated by information theory. Our first step towards this goal was conducted in [C1] where some initial results were presented comparing the achievable rates in the asymptotic block length regime to that of the finite block length regime.

6.2.2 Scheduling Policies for Machine Type Communications

The latency considered in this thesis is mainly related to the access delay and transmission delay. In general, the system delay also constitutes the queuing delay which is defined as the waiting time of a packet in the queue. While some IoT applications are event triggered and their traffic is unpredictable, other applications generate frequent and possibly periodic data that can be modelled accurately via known distributions, e.g., Poisson, geometric beta, etc. For Poisson packet arrivals, we use Markov Decision Process and linear programming in [J1] to find the optimal scheduling policy for a single user scenario. The objective of our optimization problem is to minimize the total system delay. Our results show that concatenating queued packets into larger blocks holds promising reductions in system delay when the packets are relatively short.

Appendix A

Proofs of Chapter 3

A.1 Proof of Lemma 1

As mentioned earlier, the probability that a user can be recovered after ℓ iterations of the SIC process equals to the probability that the messages received at the root of tree T_ℓ do not satisfy the SINR threshold. Starting from V nodes at depth 2ℓ, we assign each node a value 0, as none of them have been recovered yet. A C node at depth $2m-1$ will send a message v to its parent V node with the following probability:

$$Q = \sum_{d=v}^{K-1} \lambda_d \binom{d}{v} (1 - q_{m-1})^{d-v} q_{m-1}^{v},$$

which arises from the fact that a C node of degree i sends a message v to its parent if and only if exactly v out of i of its children have not been recovered yet.

On the other hand, a V node at depth $2(m-1)$ will send a message 1 to its parent if it receives messages from its children that can satisfy the SINR threshold. For a degree i V node, this happens when the messages received from its children belong to the set $\mathbf{V}^{(i+1)}$. Since the messages passed from different C nodes to the same V node parent are mutually independent (according to the tree assumption), the probability

that a degree i V node sends a message 1 to its parent can be calculated as follows:

$$Q_i = \sum_{\mathbf{v}\in\mathbf{V}^{(i+1)}} \prod_{j=1}^{i+1} \sum_{d=v_j}^{K-1} \lambda_d \binom{d}{v_j} (1-q_{\ell-1})^{d-v_j} q_{\ell-1}^{v_j}.$$

Finally, by averaging Q_i over the degree distribution $\omega(x)$, it is straightforward to arrive at the expression in (3.8).

A.2 Proof of Proposition 3.1

By substituting $\mathbf{V}_{conv}^{(i)} = \{\mathbf{v}|\exists j, v_j = 0, 1 \le j \le i\}$ instead of $\mathbf{V}^{(i)}$ in (3.8), we have:

$$q_\ell \overset{(a)}{=} \sum_{i=0}^{N-1} \omega_i \left(1 - \sum_{\mathbf{v}\in\mathbf{V}_{conv}^{(i)}} \prod_{j=1}^{i} \sum_{d=v_j}^{K-1} \lambda_d \binom{d}{v_j} (1-q_\ell)^{d-v_j} q_{\ell-1}^{v_j}\right)$$

$$\begin{aligned}
&\overset{(b)}{=} \sum_{i=0}^{N-1} \omega_i \left(\sum_{\mathbf{v}\notin\mathbf{V}_{conv}^{(i)}} \prod_{j=1}^{i} \sum_{d=v_j}^{K-1} \lambda_d \binom{d}{v_j} (1-q_\ell)^{d-v_j} q_{\ell-1}^{v_j}\right) \\
&\overset{(c)}{=} \sum_{i=0}^{N-1} \omega_i \left(\sum_{d=1}^{K-1} \sum_{v=1}^{d} \lambda_d \binom{d}{v} (1-q_\ell)^{d-v} q_{\ell-1}^{v}\right)^i \\
&\overset{(d)}{=} \sum_{i=0}^{N-1} \omega_i \left(\sum_{d=v_j}^{K-1} \lambda_d (1-(1-q_\ell)^d)\right)^i \\
&\overset{(e)}{=} \sum_{i=0}^{N-1} \omega_i \left(1 - \sum_{d=v_j}^{K-1} \lambda_d (1-q_\ell)^d\right)^i.
\end{aligned}$$

Step (a) follows from the fact that $\sum_i \omega_i = 1$. Step (b) follows from taking the conjugate of the expression in (3.5). Step (c) follows from the fact that the conjugate of $\mathbf{V}_{conv}^{(i)}$ is $\{\mathbf{v}|v_j > 0\}$, and $\{\mathbf{v}|v_j > 0\} = (\{v_j|v_j > 0\})^i$; step (d) follows from the binomial expansion of the expression given; finally, step (e) follows from $\sum_i \lambda_i = 1$.

A.3 Proof of Lemma 2

Since the sub-channels are chosen uniformly at random, the degree of each sub-channel, defined as the number of users transmitting in that sub-channel, follows the binomial distribution. Let us denote this pdf by $\Lambda(x)$. In the asymptotic case, that is for a large number of sub-channels and SUs, the distribution converges to Poisson [44] as follows:

$$\Lambda_i = e^{-\alpha}\frac{\alpha^i}{i!}, \text{ where } \quad \alpha = \frac{K_s}{N}\bar{\Omega}.$$

From (3.12), the IP constraint for PU_k was defined as: $I_p^{(i)} = \sum_{n=1}^{N_p^{(k)}} \sum_{k\in\mathcal{K}_s} a_{k,n}P_n = \sum_{n=1}^{N_p^{(k)}} u_n P_n$, where u_n is a random variable representing the number of SUs transmitting in the n^{th} sub-channel. Intuitively, the probability of having u_n SUs transmitting in a sub-channel n is the same for all $1 \leq n \leq N$, and $\Pr(I_{p,n}^{(k)} = u_n P_n)$ is simply Λ_{u_n}. Assuming equal power allocation, that is $P_n = P_o^{(k)}\ \forall\ n \in \mathcal{N}_p^{(k)}$, we can express the pdf of $I_p^{(k)}$ as follows:

$$\Pr(I_p^{(k)} = i) = \Pr(u_1P_o^{(k)} + u_2P_o^{(k)} + ... + u_{N_p^{(k)}}P_o^{(k)} = i) = \bigotimes_{n=1}^{N_p^{(k)}} \Lambda_z|_{z=\frac{i}{P_o^{(k)}}}, \forall k \in \mathcal{K}_p,$$

where $\otimes$ denotes the convolution operation. As $\Lambda(x)$ was shown to be Poisson distributed, and as the sum of Poisson distributed random variables is also a Poisson random variable, we arrive at (3.16).

Appendix B

Proofs of Chapter 4

B.1 Proof of Lemma 3

Consider an arbitrary graph $\mathcal{G}_s$ constructed with the ACK-All scheme. Let $q_i^{(s)}[\ell]$ be the probability that a Type-i OR-node remains unresolved in the ℓ^{th} iteration of the SIC process. This probability is initialized as $q_i^{(s)}[0] = \frac{|\mathcal{C}_{i,s}^{(s)}|}{|\mathcal{C}_i|}$. AND-nodes are divided into s types corresponding to the s sub-frames. A Type-i AND-node is said to have d children with a probability $\delta_d^{(i)}$. On the other hand, OR-nodes are divided into r types corresponding to the r groups. Each type of OR-node is further divided into s sub-Types corresponding to the s subsets, where the first $s-1$ sub-Types have a value of 1 (having been resolved previously). A Type-i OR-node is said to have d children with a probability of $\psi_{i,d}^{(j)}$.

A Type-i OR-node is said to be connected to a Type-j AND-node with a probability $\bar{c}_i^{(j)}$. This probability is given below as the normalized selection probability of the corresponding sub-frame by group i.

$$\bar{c}_i^{(j)} = \begin{cases} \frac{p_i^{(j)}\Delta N_j}{\sum_{j'=1}^{j} p_i^{(j')}\Delta N_{j'}} = \frac{\zeta_i^{(j)}}{\sum_{j'=1}^{j}\zeta_i^{(j')}} & \text{for sub-Types } 1, ..., j \\ 0 & \text{otherwise} \end{cases}.$$

From Figure 4.1, we find that the degrees of some AND-nodes are reduced in each

sub-frame as more of their edges are removed. Based on Equation 4.2, the probability that a Type-j AND-node is still connected to d Type-i OR-nodes in $\mathcal{G}_s$ is given as :

$$\Omega_{i,d}^{(j\to s)} = \binom{|\mathcal{C}_{i,s}^{(s)}|}{d} \left(\frac{g_i^{(j)}}{|\mathcal{C}_{i,j}^{(j)}|}\right)^d \left(1 - \frac{g_i^{(j)}}{|\mathcal{C}_{i,j}^{(j)}|}\right)^{|\mathcal{C}_{i,s}^{(s)}|-d}.$$

For $K, N \gg$, the Poisson approximation of the former equation is given as

$$\Delta^{(j\to s)}(x) = \exp\left(-\sum_{i=1}^{r} \frac{q_i^{(s)}[0]}{q_i^{(j)}[0]} g_i^{(j)}(1-x)\right), \tag{B.1}$$

Moreover, given that an edge of a Type-j AND-node has not been removed in the previous $\ell - 1$ iterations of the SIC, this edge is said to be connected to a Type-i OR-node with a probability $\bar{v}_i^{(j\to s)}[\ell]$. $\bar{v}_i^{(j\to s)}[\ell]$ is expressed below as the normalized access probability of the corresponding subset in sub-frame j.

$$\bar{v}_i^{(j\to s)} = \frac{p_i^{(j)}|\mathcal{C}_{i,s}^{(s)}|}{\sum_{i'=1}^{s} p_{i'}^{(j)}|\mathcal{C}_{i',s}^{(s)}|} = \frac{\frac{q_i^{(s)}[0]}{q_i^{(j)}[0]} g_i^{(j)}}{\sum_{i'=1}^{s} \frac{q_{i'}^{(s)}[0]}{q_{i'}^{(j)}[0]} g_{i'}^{(j)}}. \tag{B.2}$$

In general, each AND-Node at depth $2\ell - 1$ calculates its value by performing AND operation on its children at depth 2ℓ. Assume that a child of an AND-node has a value of 1 with a probability $1-q_c$. Then, the probability that an AND-node of degree d has a value of 0 is $1 - (1 - q_c)^d$. Similarly, each OR-node at depth 2ℓ calculates its value by performing OR operation on its children at depth $2\ell + 1$. Assume that a child of an OR-node has a value of 1 with a probability q_v. Then, the probability that an OR-node of degree d has a value of 0 is $(q_v)^d$. More generally, we can write

$$1 - q_{c,j} = \sum_{i=1}^{r} \bar{v}_i^{(j\to s)} \frac{q_{v,i}}{q_i^{(s)}[0]},$$

where $q_{c,j}$ is the probability that a child of an AND-node of Type-j has a value of 0. Similarly, $q_{v,i}$ is the probability that a child of an OR-node of Type-i has a value of 1. By averaging this expression over the degree distribution in Equation B.1 and the

different types of AND-nodes, we arrive at

$$\begin{aligned} q_{v,i} &= \sum_{j=1}^{s} \bar{c}_i^{(j)} \sum_d \delta_d^{(j\to s)} \left(1-(1-q_{c,j})^d\right) \\ &= 1-\sum_{j=1}^{s} \bar{c}_i^{(j)} \sum_d \delta_d^{(j\to s)} (1-q_{c,j})^d \\ &= 1-\sum_{j=1}^{s} \bar{c}_i^{(j)} \delta^{(j\to s)} (1-q_{c,j}) . \end{aligned}$$

Similarly, by averaging $q_{v,i}$ over the degree distribution in Equation 4.7, we arrive at

$$q_{v,i} = \sum_d \psi_{i,d}^{(s)} (q_{v,i})^d = \psi_i^{(s)}(q_{v,i}).$$

Finally, with a slight modification in notations, it is straightforward to arrive at the expression in Lemma 3.

B.2 Proof of Proposition 4.2

In the ACK-All scheme, devices from the group $\mathcal{C}_i$ continue to transmit in following sub-frames provided they have not been resolved in previous sub-frames. Let $a_{i,n}^{(s)}$ be a Bernoulli random variable denoting the probability that a device D from the group $\mathcal{C}_i$ has transmitted in the n^{th} time slot of the s^{th} sub-frame. Then, the average number of transmissions of a device from the group $\mathcal{C}_i$, denoted by M_i, is derived as the sum of average number of transmissions of this device in each sub-frame.

$$\begin{aligned} M_i &= \mathbb{E}\left[\sum_{s=1}^{r}\sum_{n=1}^{\Delta N_s} a_{i,n}^{(s)}\right] \\ &= \sum_{s=1}^{r} \mathbb{E}\left[\sum_{n=1}^{\Delta N_s} a_{i,n}^{(s)}\right] \\ &= \sum_{s=1}^{r} \Delta N_s \Pr\left(D\in\mathcal{C}_{i,s}^{(s)}\middle| D\in\mathcal{C}_i\right)\Pr\left(a_{i,n}^{(s)}=1\middle| D\in\mathcal{C}_{i,s}^{(s)}\right) \\ &= \sum_{s=1}^{r} \Delta N_s \frac{g_i^{(s)}}{|\mathcal{C}_{i,s}^{(s)}|}\frac{\left|C_{i,s}^{(s)}\right|}{|\mathcal{C}_i|} = \sum_{s=1}^{r} \Delta N_s \frac{g_i^{(s)}}{|\mathcal{C}_i|}. \end{aligned}$$

B.3 Proof of Proposition 4.3

Following on the analogy between our RA schemes and codes on graph, the system load is analogous to the code rate. The code rate is the ratio of the number of information symbols to the number of coded symbols. The channel capacity dictates the maximum achievable code rate for a given channel SNR at which reliable transmission is possible. Similarly, the bound $L^*(\epsilon)$ dictates the maximum system load for a given target error probability ϵ at which a reliable communication is possible. While the channel capacity can be calculated from the Shannon bound, $L^*(\epsilon)$ can be calculated from the AND-OR tree equations given in Proposition 4.1. For example, in Figure 4.6, the achievable error probabilities are shown for different loads and different values of g for the case of $r = 1$. For $\epsilon = 0.02$, the system load is said to be upper bounded by $1/1.2$, i.e., there exists a $g > 0$ that can guarantee an average probability of device resolution error of 0.02 for every K/N less than $1/1.2$. It is straightforward to generalize this concept for $r > 1$ for the ACK-Group scheme. Moreover, as the ACK-Group scheme is a special case of the ACK-All scheme, we can say that if there exists a matrix $\mathbf{G}$ that can satisfy the latency requirements of all the groups in the ACK-Group scheme, then, there exists a matrix $\mathbf{G}$ that can satisfy these requirements in the ACK-All scheme as well. The argument of L^* in Equation 4.11 takes into consideration that some of the devices of $\mathcal{C}_i$ have been resolved in the previous $i-1$ sub-frames. Therefore, the error probability is scaled according to the number of unresolved devices. For the ACK-Group scheme, $\epsilon_i^{(s)} = 1$ for $i \leq s$.

Appendix C

Proofs of Chapter 5

C.1 Proof of Corollary 1

For a device uniformly distributed in an annulus of minimum radius $d_{\min}$ and maximum radius $d_{\max}$, the CDF of the distance d with respect to the origin follows the truncated Pareto distribution below

$$F_D(d) = \frac{d^2 - d_{\min}^2}{d_{\max}^2 - d_{\min}^2}.$$

For $X = d^{-\alpha}$,

$$\begin{aligned} F_X(x) &= \Pr(d^{-\alpha} < x) \\ &= \Pr(d > x^{-\frac{1}{\alpha}}) \\ &= 1 - \frac{x^{-\frac{2}{\alpha}} - d_{\min}^2}{d_{\max}^2 - d_{\min}^2}. \end{aligned}$$

For Rayleigh fading, we can write

$$F_P(p) = \int_0^\infty \Pr(vd^{-\alpha} < p|v)\Pr(v)dv$$

$$= \int_0^\infty \left(1 - \frac{\left(\frac{p}{v}\right)^{-\frac{2}{\alpha}} - d_{\min}^2}{d_{\max}^2 - d_{\min}^2}\right) e^{-v} \mathrm{d}v.$$

$$= 1 - \int_0^\infty \frac{\left(\frac{p}{v}\right)^{-\frac{2}{\alpha}} - d_{\min}^2}{d_{\max}^2 - d_{\min}^2} e^{-v} \mathrm{d}v.$$

$$= 1 - \frac{p^{-\frac{2}{\alpha}}}{d_{\max}^2 - d_{\min}^2} \Gamma\left[\frac{2}{\alpha} + 1\right] + \frac{d_{\min}^2}{d_{\max}^2 - d_{\min}^2}.$$

C.2 Proof of Lemma 4

Consider a vector $\mathbf{N} = \{N_i\}_{1 \leq i \leq L}$, where $N_i \in \mathbb{N}$ are independent and identically distributed Poisson distributed random variables with average λ. For that, we define the following events:

$$\begin{aligned} A_i &: && \text{Event of } N_i = 1. \\ B &: && \text{Event of } N_i \neq 1, \forall i. \\ S_n &: && \text{Event of } \sum_{i=1}^{L} N_i = n. \end{aligned}$$

Following from the definition of Poisson distributions, we have

$$\begin{aligned} &\Pr(A_i) = \Pr(A) = \lambda e^{-\lambda}, \quad \forall i, \\ &\Pr(B) = (1 - \Pr(A))^L = \left(1 - \lambda e^{-\lambda}\right)^L, \text{ and} \\ &\Pr\left(S_n\right) = \frac{(L\lambda)^n e^{-L\lambda}}{n!}. \end{aligned}$$

The last equation follows from the fact that the sum of any c elements of $\mathbf{N}$ is a Poisson distributed random variable with average $c\lambda$.

We want to calculate the probability that the sum of these elements is equal to some

positive integer n given that $N_i \neq 1$ for $1 \leq i \leq L$. This can be expressed as

$$\Pr\left(S_n|B\right) = \frac{\Pr\left(B|S_n\right)\Pr\left(S_n\right)}{\Pr\left(B\right)} = \frac{\left(1-\Pr\left(\bar{B}|S_n\right)\right)\Pr\left(S_n\right)}{1-\Pr\left(\bar{B}\right)}.$$

Moreover, we have

$$\Pr\left(A_1 \cap A_2 \cap ...A_c \middle| S_n\right) \quad = \quad \Pr\left(A\right)^c \frac{\left((L-c)\lambda\right)^{n-c} e^{-(L-c)\lambda}}{(n-c)!},$$

for $c \leq \min\{L-1, n\}$; it is zero otherwise. Given S_n, the probability that at least one of the elements of vector $\mathbf{N}$ is equal to 1 can be expressed as:

$$\begin{aligned}
\Pr\left(\bar{B}|S_n\right) &= \Pr\left(\bigcup_{i=1}^{L} A_i \middle| S_n\right) \\
&= \sum_{c=1}^{L}(-1)^{c+1} \sum_{\substack{i_1,\ldots,i_c: \\ 1\leq i_1<\ldots<i_c\leq L}} \Pr\left(A_{i_1} \cap \ldots \cap A_{i_c}|S_n\right) \\
&= L\Pr(A)\frac{\left((L-1)\lambda\right)^{n-1} e^{-(L-1)\lambda}}{(n-1)!} - \\
&\binom{L}{2} \Pr\left(A\right)^2 \frac{\left((L-2)\lambda\right)^{n-2} e^{-(L-2)\lambda}}{(n-2)!} + \ldots + \\
&(-1)^{\theta}\binom{L}{\theta} \Pr\left(A\right)^{\theta} \frac{\left((L-\theta)\lambda\right)^{n-\theta} e^{-(L-\theta)\lambda}}{(n-\theta)!} \\
&= \sum_{c=1}^{\theta}(-1)^{c+1}\binom{L}{c}\left(\lambda e^{-\lambda}\right)^c \frac{\left((L-c)\lambda\right)^{n-c} e^{-(L-c)\lambda}}{(n-c)!},
\end{aligned}$$

where $\theta = \min\{L-1, n\}$. Then, it is straightforward to arrive at Lemma 4 by calculating $\frac{\left(1-\Pr\left(\bar{B}|S_n\right)\right)\Pr(S_n)}{1-\Pr\left(\bar{B}\right)}$ and substituting L by $L - L_s$ and λ by $\frac{\lambda}{L}$.

C.3 Proof of Lemma 5

$$\epsilon_{\mathrm{SJD}} \overset{(a)}{=} 1 - \mathbb{E}_{L_s}\left[\int_0^\infty \max_{v}\left\{v, R \leq \min_{0\leq u\leq 1} \frac{1}{uvL_s}\log\left(1+\frac{g(v)-g(v-uv)}{g(1)-g(v)+\frac{x+\sigma^2}{L_s}}\right)\right\} f_I(x)\mathrm{d}I\right]$$
$$\overset{(b)}{=} 1 - \mathbb{E}_{L_s}\left[\int_0^1 v f_I(I^*_{\mathrm{SJD}}(v))\mathrm{d}v\right].$$

$$\mathbb{E}_I\left[\max_{0\leq v\leq 1}\min_{0<u\leq 1}\frac{1}{u}\log\left(1+\frac{g(v)-g(v-uv)}{g(1)-g(v)+\frac{I+\sigma^2}{L_s}}\right)\right] \overset{(c)}{=}$$
$$\mathbb{E}_I\left[\max_{0\leq v\leq 1}\min_{0<u\leq 1}\frac{1}{u}\int_0^\infty \frac{e^{-z}}{z}\left(1-e^{\frac{-z(g(v)-g(v-uv))}{g(1)-g(v)+\frac{I+\sigma^2}{L_s}}}\right)\mathrm{d}z\right] \overset{(d)}{=}$$
$$\mathbb{E}_I\left[\max_{0\leq v\leq 1}\min_{0<u\leq 1}\frac{1}{u}\int_0^\infty \frac{e^{-s((g(1)-g(v))L_s+\sigma^2)}}{s}e^{-sI}\left(1-e^{-s(g(v)-g(v-uv))L_s}\right)\mathrm{d}s\right] \overset{(e)}{=}$$
$$\max_{0\leq v\leq 1}\min_{0<u\leq 1}\frac{1}{u}\int_0^\infty \frac{e^{-s((g(1)-g(v))L_s+\sigma^2)}}{s}\mathbb{E}_I\left[e^{-sI}\right]\left(1-e^{-s(g(v)-g(v-uv))L_s}\right)\mathrm{d}s \overset{(f)}{=}$$
$$\max_{0\leq v\leq 1}\min_{0<u\leq 1}\frac{1}{u}\int_0^\infty \frac{e^{-s((g(1)-g(v))L_s+\sigma^2)}}{s}\psi_I(-s)\left(1-e^{-s(g(v)-g(v-uv))L_s}\right)\mathrm{d}s.$$

From Equation 5.19 and Equation 5.22, the outage probability converges to a deterministic value given below as

$$\epsilon_{\mathrm{SJD}}(L_s, I, \hat{\mathbf{P}}) = 1 - \max_{0\leq v\leq L_s}\left\{v, R \leq \min_{0<u\leq 1}\frac{1}{uvL_s}\log\left(1+\frac{g(v)-g(u-uv)}{g(1)-g(v)+\frac{I+\sigma^2}{L_s}}\right)\right\},$$

where $u := \frac{i}{\ell}$ and $v := \frac{\ell}{L_s}$. As we are considering a massive access setting, the average number of singleton layers $\lambda e^{-\frac{\lambda}{L}}$ is taken to be asymptotically large. Moreover, as the ordered received powers converge to deterministic values under massive access, we can see that expression above for the outage probability for a given number of singleton layers depends only on the interference power. Based on this, in step (a), we average over L_s and I. Then, in step (b), we decompose the integral such that

$I^*_{\mathrm{SJD}}(\frac{\ell}{L_s})$ denotes the maximum interference power for which ℓ out of L_s layers can be decoded correctly. It is expressed in Equation 5.23. Finally, we arrive at Lemma 5 by applying the inversion theorem in (5.14).

C.4 Proof of Lemma 6

The average system throughput can be evaluated by averaging the expression in (5.25) over L_s and I. Step (c) follows from averaging over the latter and using the following identity [144]:

$$\log(1+x) = \int_0^\infty \frac{e^{-z}}{z}\left(1 - e^{-xz}\right).$$

Step (d) follows from a change of variable: $z = s\,(I + \sigma^2)$, and step (e) follows from interchanging the integration and the expectation. Finally, step (f) follows from the definition of CFs.

List of References

[1] A. Al-Fuqaha, M. Guizani, M. Mohammadi, M. Aledhari, and M. Ayyash, "Internet of things: A survey on enabling technologies, protocols, and applications," *IEEE Communications Surveys & Tutorials*, vol. 17, no. 4, pp. 2347–2376, 2015.

[2] R. Van Kranenburg, "A critique of ambient technology and the all-seeing network of RFID," 2008.

[3] D. Giusto, "A. lera, g. morabito, l. atzori (eds.) the internet of things," 2010.

[4] L. Atzori, A. Iera, and G. Morabito, "The internet of things: A survey," *Computer networks*, vol. 54, no. 15, pp. 2787–2805, 2010.

[5] "Institute of Electrical and Electronics Engineers (IEEE): The Internet of Things (IoT)," Special Report, Mar. 2014.

[6] "International Telecommunication Union (ITU): The Internet of Things (IoT)," Internet Report, Nov. 2005.

[7] J. Gubbi, R. Buyya, S. Marusic, and M. Palaniswami, "Internet of things (IoT): A vision, architectural elements, and future directions," *Future generation computer systems*, vol. 29, no. 7, pp. 1645–1660, 2013.

[8] A. Zanella, N. Bui, A. Castellani, L. Vangelista, and M. Zorzi, "Internet of things for smart cities," *IEEE Internet of Things Journal*, vol. 1, no. 1, pp. 22–32, 2014.

[9] L. Atzori, A. Iera, and G. Morabito, "From" smart objects" to" social objects": The next evolutionary step of the internet of things," *IEEE Communications Magazine*, vol. 52, no. 1, pp. 97–105, 2014.

[10] S. C. Mukhopadhyay and N. Suryadevara, "Internet of things: Challenges and opportunities," in *Internet of Things*. Springer, 2014, pp. 1–17.

[11] J. A. Stankovic, "Research directions for the internet of things," *IEEE Internet of Things Journal*, vol. 1, no. 1, pp. 3–9, 2014.

[12] L. Da Xu, W. He, and S. Li, "Internet of things in industries: A survey," *IEEE Transactions on Industrial Informatics*, vol. 10, no. 4, pp. 2233–2243, 2014.

[13] M. A. Lema, A. Laya, T. Mahmoodi, M. Cuevas, J. Sachs, J. Markendahl, and M. Dohler, "Business case and technology analysis for 5G low latency applications," *IEEE Access*, vol. 5, pp. 5917–5935, 2017.

[14] K. Ashton, "That 'internet of things' thing," *RFiD Journal*, vol. 22, no. 7, 2011.

[15] S. Fan, L. Zhang, W. Feng, W. Zhang, and Y. Ren, "Optimization-based design of wireless link scheduling with physical interference model," *IEEE Transactions on Vehicular Technology*, vol. 61, no. 8, pp. 3705–3717, Oct 2012.

[16] C. E. Shannon, "Probability of error for optimal codes in a gaussian channel," *Bell Labs Technical Journal*, vol. 38, no. 3, pp. 611–656, 1959.

[17] G. Durisi, T. Koch, and P. Popovski, "Toward massive, ultrareliable, and low-latency wireless communication with short packets," *Proceedings of the IEEE*, vol. 104, no. 9, pp. 1711–1726, 2016.

[18] V. Strassen, "Asymptotische abschatzungen in shannon's informationstheorie," in *Transaction of the 3rd Prague Conference on Information Theory*. Prague, 1962, pp. 689–723.

[19] Y. Polyanskiy, H. V. Poor, and S. Verdú, "Channel coding rate in the finite blocklength regime," *IEEE Transactions on Information Theory*, vol. 56, no. 5, pp. 2307–2359, 2010.

[20] M. Hayashi, "Information spectrum approach to second-order coding rate in channel coding," *IEEE Transactions on Information Theory*, vol. 55, no. 11, pp. 4947–4966, 2009.

[21] V. Y. F. Tan and M. Tomamichel, "The third-order term in the normal approximation for the awgn channel," *IEEE Transactions on Information Theory*, vol. 61, no. 5, pp. 2430–2438, 2015.

[22] R. Gallager, "A simple derivation of the coding theorem and some applications," *IEEE Transactions on Information Theory*, vol. 11, no. 1, pp. 3–18, 1965.

[23] R. G. Gallager, *Information theory and reliable communication*. Springer, 1968, vol. 2.

[24] L. Weng, S. S. Pradhan, and A. Anastasopoulos, "Error exponent regions for gaussian broadcast and multiple-access channels," *IEEE Transactions on Information Theory*, vol. 54, no. 7, pp. 2919–2942, 2008.

[25] E. Haim, Y. Kochman, and U. Erez, "Improving the MAC error exponent using distributed structure," in *Information Theory Proceedings (ISIT), 2011 IEEE International Symposium on*. IEEE, 2011, pp. 1003–1007.

[26] R. Ahlswede, "Multi-way communication channels," in *Second International Symposium on Information Theory: Tsahkadsor, Armenia, USSR, Sept. 2-8, 1971*, 1973.

[27] H. Liao, "A coding theorem for multiple access communications," in *Proc. Int. Symp. Information Theory, Asilomar, CA, 1972*, 1972.

[28] T. M. Cover, "Some advances in broadcast channels," *Advances in communication systems*, vol. 4, pp. 229–260, 1975.

[29] A. Wyner, "Recent results in the shannon theory," *IEEE Transactions on information Theory*, vol. 20, no. 1, pp. 2–10, 1974.

[30] S. V. Hanly, "Information capacity of radio networks," Ph.D. dissertation, University of Cambridge, 1993.

[31] E. MolavianJazi and J. N. Laneman, "A finite-blocklength perspective on gaussian multi-access channels," *arXiv preprint arXiv:1309.2343*, 2013.

[32] ——, "On the second-order cost of TDMA for gaussian multiple access," in *2014 IEEE International Symposium on Information Theory*. IEEE, 2014, pp. 266–270.

[33] N. Abramson, "The ALOHA system: another alternative for computer communications," in *Proceedings of the November 17-19, 1970, fall joint computer conference*. ACM, 1970, pp. 281–285.

[34] L. G. Roberts, "ALOHA packet system with and without slots and capture," *ACM SIGCOMM Computer Communication Review*, vol. 5, no. 2, pp. 28–42, 1975.

[35] N. Abramson, "The throughput of packet broadcasting channels," *IEEE Transactions on Communications*, vol. 25, no. 1, pp. 117–128, 1977.

[36] W. Crowther, "A system for broadcast communication: Reservation-ALOHA," *Proc. IEEE HICSS, Jan. 1973*, pp. 596–603, 1973.

[37] E. Casini, R. De Gaudenzi, and O. D. R. Herrero, "Contention resolution diversity slotted ALOHA (CRDSA): An enhanced random access schemefor satellite access packet networks," *IEEE Transactions on Wireless Communications*, vol. 6, no. 4, 2007.

[38] Y. Yu and G. Giannakis, "High-throughput random access using successive interference cancellation in a tree algorithm," *IEEE Transactions on Information Theory*, vol. 53, no. 12, pp. 4628–4639, Dec 2007.

[39] S. Gollakota and D. Katabi, "Zigzag decoding: combating hidden terminals in wireless networks," in *Proceedings of the ACM SIGCOMM conference on Data communication*, vol. 38, no. 4. ACM, 2008.

[40] A. S. Tehrani, A. G. Dimakis, and M. J. Neely, "Sigsag: Iterative detection through soft message-passing," *IEEE Journal of Selected Topics in Signal Processing*, vol. 5, no. 8, pp. 1512–1523, 2011.

[41] G. Liva, "Graph-based analysis and optimization of contention resolution diversity slotted ALOHA," *IEEE Transactions on Communications*, vol. 59, no. 2, pp. 477–487, 2011.

[42] T. J. Richardson, M. A. Shokrollahi, and R. L. Urbanke, "Design of capacity-approaching irregular low-density parity-check codes," *IEEE Transactions on Information Theory*, vol. 47, no. 2, pp. 619–637, 2001.

[43] M. Luby, "LT codes," in *The 43rd annual IEEE Symposium on foundations in Computer Science*, 2002, pp. 271–280.

[44] A. Shokrollahi, "Raptor codes," *IEEE Transactions on Information Theory*, vol. 52, no. 6, pp. 2551–2567, 2006.

[45] M. Luby, M. Mitzenmacher, and M. A. Shokrollahi, "Analysis of random processes via And-Or tree evaluation," in *SODA*, vol. 98, 1998, pp. 364–373.

[46] C. Stefanovic, P. Popovski, and D. Vukobratovic, "Frameless ALOHA protocol for wireless networks," *IEEE Communications Letters*, vol. 16, no. 12, pp. 2087–2090, 2012.

[47] C. Stefanovic and P. Popovski, "ALOHA random access that operates as a rateless code," *IEEE Transactions on Communications*, vol. 61, no. 11, pp. 4653–4662, 2013.

[48] C. Stefanovic, M. Momoda, and P. Popovski, "Exploiting capture effect in frameless ALOHA for massive wireless random access," in *IEEE Wireless Communications and Networking Conference (WCNC), 2014*, 2014, Conference Proceedings, pp. 1762–1767.

[49] A. Zanella, M. Zorzi, A. F. dos Santos, P. Popovski, N. Pratas, C. Stefanovic, A. Dekorsy, C. Bockelmann, B. Busropan, and T. Norp, "M2M massive wireless access: challenges, research issues, and ways forward," in *IEEE Globecom Workshops (GC Wkshps), 2013*. IEEE, 2013, Conference Proceedings, pp. 151–156.

[50] "Third Generation Partnership program (3GPP): Service Requirements for Machine-Type Communications (MTC); Stage 1," Technical Specification 22.368, V.14.0.0, Mar. 2017.

[51] A. Laya, L. Alonso, and J. Alonso-Zarate, "Is the random access channel of LTE and LTE-A suitable for M2M communications? a survey of alternatives," *IEEE Communications Surveys & Tutorials*, vol. 16, no. 1, pp. 4–16, 2014.

[52] S. Choi, W. Lee, D. Kim, K.-J. Park, S. Choi, and K.-Y. Han, "Automatic configuration of random access channel parameters in LTE systems," in *2011 IFIP Wireless Days (WD)*. IEEE, 2011, pp. 1–6.

[53] A. Lo, Y. W. Law, M. Jacobsson, and M. Kucharzak, "Enhanced lte-advanced random-access mechanism for massive machine-to-machine (M2M) communications," in *27th World Wireless Research Forum (WWRF) Meeting*. WWRF27-WG4-08„ 2011, pp. 1–5.

[54] D. T. Wiriaatmadja and K. W. Choi, "Hybrid random access and data transmission protocol for machine-to-machine communications in cellular networks," *IEEE Transactions on Wireless Communications*, vol. 14, no. 1, pp. 33–46, 2015.

[55] K. S. Ko, M. J. Kim, K. Y. Bae, D. K. Sung, J. H. Kim, and J. Y. Ahn, "A novel random access for fixed-location machine-to-machine communications in OFDMA based systems," *IEEE Communications Letters*, vol. 16, no. 9, pp. 1428–1431, 2012.

[56] Z. Wang and V. Wong, "Optimal access class barring for stationary machine type communication devices with timing advance information," *IEEE Transactions on Wireless Communications*, 2015.

[57] T. M. Cover, "Broadcast channels," *IEEE Transactions on Information Theory*, vol. 18, no. 1, pp. 2–14, Jan 1972.

[58] A. Benjebbour, Y. Saito, Y. Kishiyama, A. Li, A. Harada, and T. Nakamura, "Concept and practical considerations of non-orthogonal multiple access (NOMA) for future radio access," in *2013 International Symposium on Intelligent Signal Processing and Communication Systems*, Nov 2013, pp. 770–774.

[59] S. Han, I. Chih-Lin, Z. Xu, and Q. Sun, "Energy efficiency and spectrum efficiency co-design: From NOMA to network NOMA," *E-LETTER*, 2014.

[60] W. Shin, M. Vaezi, B. Lee, D. J. Love, J. Lee, and H. V. Poor, "Coordinated beamforming for multi-cell mimo-noma," *IEEE Communications Letters*, vol. 21, no. 1, pp. 84–87, 2017.

[61] W. Shin, M. Vaezi, J. Lee, and H. V. Poor, "On the number of users served in mimo-noma cellular networks," in *Wireless Communication Systems (ISWCS), 2016 International Symposium on*. IEEE, 2016, pp. 638–642.

[62] Y. Endo, Y. Kishiyama, and K. Higuchi, "Uplink non-orthogonal access with MMSE-SIC in the presence of inter-cell interference," in *IEEE International Symposium on Wireless Communication Systems (ISWCS)*, 2012, pp. 261–265.

[63] N. Zhang, J. Wang, G. Kang, and Y. Liu, "Uplink nonorthogonal multiple access in 5G systems," *IEEE Communications Letters*, vol. 20, no. 3, pp. 458–461, 2016.

[64] X. Chen, Z. Zhang, C. Zhong, and D. W. K. Ng, "Exploiting multiple-antenna techniques for non-orthogonal multiple access," *arXiv preprint arXiv:1706.00635*, 2017.

[65] S. Qureshi and S. A. Hassan, "Mimo uplink noma with successive bandwidth division," in *Wireless Communications and Networking Conference Workshops (WCNCW), 2016 IEEE*. IEEE, 2016, pp. 481–486.

[66] A. Li, A. Benjebbour, X. Chen, H. Jiang, and H. Kayama, "Uplink non-orthogonal multiple access (noma) with single-carrier frequency division multiple access (SC-FDMA) for 5G systems," *IEICE Transactions on Communications*, vol. 98, no. 8, pp. 1426–1435, 2015.

[67] X. Chen, A. Benjebbour, A. Li, and A. Harada, "Multi-user proportional fair scheduling for uplink non-orthogonal multiple access (NOMA)," in *IEEE 79th Vehicular Technology Conference (VTC Spring)*. IEEE, 2014, pp. 1–5.

[68] M. A. Sedaghat and R. R. Müller, "On user pairing in noma uplink," *arXiv preprint arXiv:1707.01846*, 2017.

[69] T. Takeda and K. Higuchi, "Enhanced user fairness using non-orthogonal access with SIC in cellular uplink," in *IEEE Vehicular Technology Conference (VTC Fall)*, 2011, pp. 1–5.

[70] M. S. Ali, H. Tabassum, and E. Hossain, "Dynamic user clustering and power allocation for uplink and downlink non-orthogonal multiple access (noma) systems," *IEEE Access*, vol. 4, pp. 6325–6343, 2016.

[71] K. Higuchi and A. Benjebbour, "Non-orthogonal multiple access (noma) with successive interference cancellation for future radio access," *IEICE Transactions on Communications*, vol. 98, no. 3, pp. 403–414, 2015.

[72] F. Al Rabee, K. Davaslioglu, and R. Gitlin, "The optimum received power levels of uplink non-orthogonal multiple access (NOMA) signals," in *IEEE*

18th Wireless and Microwave Technology Conference (WAMICON). IEEE, 2017, pp. 1–4.

[73] R. Hoshyar, R. Razavi, and M. Al-Imari, "LDS-OFDM an efficient multiple access technique," in *IEEE 71st Vehicular Technology Conference (VTC-Spring)*. IEEE, 2010, pp. 1–5.

[74] H. Nikopour and H. Baligh, "Sparse code multiple access," in *IEEE 24th Annual International Symposium on Personal, Indoor, and Mobile Radio Communications (PIMRC)*. IEEE, 2013, pp. 332–336.

[75] S. Zhang, X. Xu, L. Lu, Y. Wu, G. He, and Y. Chen, "Sparse code multiple access: An energy efficient uplink approach for 5G wireless systems," in *IEEE Global Communications Conference*. IEEE, 2014, pp. 4782–4787.

[76] M. Taherzadeh, H. Nikopour, A. Bayesteh, and H. Baligh, "SCMA codebook design," in *IEEE 80th Vehicular Technology Conference (VTC2014-Fall)*. IEEE, 2014, pp. 1–5.

[77] S. Chen, B. Ren, Q. Gao, S. Kang, S. Sun, and K. Niu, "Pattern division multiple access—a novel nonorthogonal multiple access for fifth-generation radio networks," *IEEE Transactions on Vehicular Technology*, vol. 66, no. 4, pp. 3185–3196, 2017.

[78] T. Abe and T. Matsumoto, "Space-time turbo equalization in frequency-selective mimo channels," *IEEE Transactions on Vehicular Technology*, vol. 52, no. 3, pp. 469–475, 2003.

[79] J. Zhang, S. Chen, X. Mu, and L. Hanzo, "Evolutionary-algorithm-assisted joint channel estimation and turbo multiuser detection/decoding for ofdm/sdma," *IEEE Transactions on Vehicular Technology*, vol. 63, no. 3, pp. 1204–1222, 2014.

[80] L. Hanzo, M. Münster, B. Choi, and T. Keller, *OFDM and MC-CDMA for broadband multi-user communications, WLANs and broadcasting*. John Wiley & Sons, 2005.

[81] M. Y. Alias, S. Chen, and L. Hanzo, "Multiple-antenna-aided ofdm employing genetic-algorithm-assisted minimum bit error rate multiuser detection," *IEEE Transactions on Vehicular Technology*, vol. 54, no. 5, pp. 1713–1721, 2005.

[82] "Third Generation Partnership program (3GPP), technical specification group radio access network, study on downlink multiuser superposition transmission (MUST) for LTE," Technical Specification 36.859, V.13.0.0, Dec. 2015.

[83] A.-H. Mohsenian-Rad, V. W. Wong, and R. Schober, "Optimal sinr-based random access," in *Proceedings of IEEE INFOCOM*, 2010, pp. 1–9.

[84] D. N. M. Dang, C. S. Hong, S. Lee, and J. Lee, "A sinr-based MAC protocol for wireless ad hoc networks," *IEEE Communications Letters*, vol. 16, no. 12, pp. 2016–2019, December 2012.

[85] M. H. Cheung and V. Wong, "Interference pricing for sinr-based random access game," *IEEE Transactions on Wireless Communications*, vol. 12, no. 5, pp. 2292–2301, May 2013.

[86] M. H. Cheung, V. W. Wong, and R. Schober, "Sinr-based random access for cognitive radio: Distributed algorithm and coalitional game," *IEEE Transactions on Wireless Communications*, vol. 10, no. 11, pp. 3887–3897, 2011.

[87] O. Ben Sik Ali, C. Cardinal, and F. Gagnon, "On the performance of interference cancellation in wireless ad hoc networks," *IEEE Transactions on Communications*, vol. 58, no. 2, pp. 433–437, February 2010.

[88] Y.-C. Liang, K.-C. Chen, Y. Li, and P. Mahonen, "Cognitive radio networking and communications: an overview," *IEEE Transactions on Vehicular Technology*, vol. 60, no. 7, pp. 3386–3407, Sept. 2011.

[89] A. De Domenico, E. Strinati, and M. Di Benedetto, "A survey on MAC strategies for cognitive radio networks," *IEEE Commununications Surveys & Tutorials*, vol. 14, no. 1, pp. 21–44, 2012.

[90] S. Barman Roy, S. N. Merchant, and A. Madhukumar, "Random transmission in cognitive uplink network," in *International Conference on Mobile Services, Resources, and Users (MOBILITY)*, 2013, pp. 94–98.

[91] J. Lim, H. G. Myung, K. Oh, and D. J. Goodman, "Channel-dependent scheduling of uplink single carrier fdma systems," in *Vehicular technology conference, 2006. VTC-2006 Fall. 2006 IEEE 64th.* IEEE, 2006, pp. 1–5.

[92] F. F. Digham, "Joint power and channel allocation for cognitive radios," in *Wireless Communications and Networking Conference (WCNC).* IEEE, 2008, pp. 882–887.

[93] R. Storn and K. Price, "Differential evolution–a simple and efficient heuristic for global optimization over continuous spaces," *Journal of Global Optimization*, vol. 11, no. 4, pp. 341–359, 1997.

[94] E. Paolini, G. Liva, and M. Chiani, "High throughput random access via codes on graphs: Coded slotted ALOHA," in *IEEE International Conference on Communications (ICC)*, 2011, pp. 1–6.

[95] G. Liva, E. Paolini, M. Lentmaier, and M. Chiani, "Spatially-coupled random access on graphs," in *IEEE International Symposium on Information Theory Proceedings (ISIT)*, 2012, pp. 478–482.

[96] N. Hansen and A. Ostermeier, "Adapting arbitrary normal mutation distributions in evolution strategies: The covariance matrix adaptation," in *Proceedings of IEEE International Conference on Evolutionary Computation*, 1996, pp. 312–317.

[97] R. Zhang, "On peak versus average interference power constraints for protecting primary users in cognitive radio networks," *IEEE Transactions on Wireless Communications*, vol. 8, no. 4, pp. 2112–2120, Apr. 2009.

[98] H. Xu and B. Li, "Efficient resource allocation with flexible channel cooperation in ofdma cognitive radio networks," in *Proceedings of IEEE INFOCOM*, 2010, pp. 1–9.

[99] A. Biral, M. Centenaro, A. Zanella, L. Vangelista, and M. Zorzi, "The challenges of M2M massive access in wireless cellular networks," *Digital Communications and Networks*, vol. 1, no. 1, pp. 1–19, 2015.

[100] S.-Y. Lien, K.-C. Chen, and Y. Lin, "Toward ubiquitous massive accesses in 3GPP machine-to-machine communications," *IEEE Communications Magazine*, vol. 49, no. 4, pp. 66–74, 2011.

[101] P. Si, J. Yang, S. Chen, and H. Xi, "Adaptive massive access management for QoS guarantees in M2M communications," *IEEE Transactions on Vehicular Technology*, vol. 64, no. 7, pp. 3152–3166, 2015.

[102] M. Shirvanimoghaddam, Y. Li, M. Dohler, B. Vucetic, and S. Feng, "Probabilistic rateless multiple access for machine-to-machine communication," *IEEE Transactions on Wireless Communications*, vol. 14, no. 12, pp. 6815–6826, 2015.

[103] H. S. Dhillon, H. Huang, H. Viswanathan, and R. A. Valenzuela, "Fundamentals of throughput maximization with random arrivals for m2m communications," *IEEE Transactions on Communications*, vol. 62, no. 11, pp. 4094–4109, 2014.

[104] Y. Zhang, "Tree-based resource allocation for periodic cellular m2m communications," *IEEE Wireless Communications Letters*, vol. 3, no. 6, pp. 621–624, 2014.

[105] C. Kahn and H. Viswanathan, "Connectionless access for mobile cellular networks," *IEEE Communications Magazine*, vol. 53, no. 9, pp. 26–31, 2015.

[106] C. Stefanovic and P. Popovski, "Coded slotted aloha with varying packet loss rate across users," in *IEEE Global Conference on Signal and Information Processing (GlobalSIP)*. IEEE, 2013, pp. 787–790.

[107] D. Sejdinovic, R. Piechocki, A. Doufexi, and M. Ismail, "Decentralised distributed fountain coding: asymptotic analysis and design," *IEEE Communications Letters*, vol. 14, no. 1, pp. 42–44, 2010.

[108] L. Toni and P. Frossard, "Prioritized random MAC optimization via graph-based analysis," *IEEE Transactions on Communications*, vol. 63, no. 12, pp. 5002–5013, 2015.

[109] M. Ivanov, F. Brännström, A. G. i Amat, and G. Liva, "Unequal error protection in coded slotted aloha," *IEEE Wireless Communications Letters*, vol. 5, no. 5, pp. 536–539, Oct 2016.

[110] P. Popovski, "Ultra-reliable communication in 5G wireless systems," in *5G for Ubiquitous Connectivity (5GU), 2014 1st International Conference on*. IEEE, 2014, pp. 146–151.

[111] B. Wang, Y. Wu, F. Han, Y.-H. Yang, and K. R. Liu, "Green wireless communications: A time-reversal paradigm," *IEEE Journal on Selected Areas in Communications*, vol. 29, no. 8, pp. 1698–1710, 2011.

[112] Y. Chen, F. Han, Y.-H. Yang, H. Ma, Y. Han, C. Jiang, H.-Q. Lai, D. Claffey, Z. Safar, and K. R. Liu, "Time-reversal wireless paradigm for green internet of things: An overview," *IEEE Internet of Things Journal*, vol. 1, no. 1, pp. 81–98, 2014.

[113] N. Rahnavard, B. N. Vellambi, and F. Fekri, "Rateless codes with unequal error protection property," *IEEE Transactions on Information Theory*, vol. 53, no. 4, pp. 1521–1532, 2007.

[114] M. Luby, M. Mitzenmacher, A. Shokrollah, and D. Spielman, "Analysis of low density codes and improved designs using irregular graphs," in *Proceedings of the* 30^{th} *annual ACM Symposium on Theory of Computing*. ACM, 1998, pp. 249–258.

[115] S.-Y. Chung, T. J. Richardson, and R. L. Urbanke, "Analysis of sum-product decoding of low-density parity-check codes using a Gaussian approximation," *IEEE Transactions on Information Theory*, vol. 47, no. 2, pp. 657–670, 2001.

[116] J. H. Sørensen, P. Popovski, and J. Østergaard, "Feedback in LT codes for prioritized and non-prioritized data," in *IEEE Vehicular Technology Conference (VTC Fall)*. IEEE, 2012, pp. 1–5.

[117] J.-P. Cheng, C.-h. Lee, and T.-M. Lin, "Prioritized random access with dynamic access barring for RAN overload in 3GPP LTE-A networks," in *2011 IEEE GLOBECOM Workshops (GC Wkshps)*. IEEE, 2011, pp. 368–372.

[118] Č. Stefanović, K. F. Trilingsgaard, N. K. Pratas, and P. Popovski, "Joint estimation and contention-resolution protocol for wireless random access," in *2013 IEEE International Conference on Communications (ICC)*, June 2013, pp. 3382–3387.

[119] S. Y. Lien, S. C. Hung, K. C. Chen, and Y. C. Liang, "Ultra-low-latency ubiquitous connections in heterogeneous cloud radio access networks," *IEEE Wireless Communications*, vol. 22, no. 3, pp. 22–31, June 2015.

[120] E. MolavianJazi and J. N. Laneman, "A random coding approach to gaussian multiple access channels with finite blocklength," in *IEEE 50th Annual Allerton Conference on Communication, Control, and Computing (Allerton)*, 2012, pp. 286–293.

[121] L. Dai, B. Wang, Y. Yuan, S. Han, I. Chih-Lin, and Z. Wang, "Non-orthogonal multiple access for 5G: solutions, challenges, opportunities, and future research trends," *IEEE Communications Magazine*, vol. 53, no. 9, pp. 74–81, 2015.

[122] C. Bockelmann, N. Pratas, H. Nikopour, K. Au, T. Svensson, C. Stefanovic, P. Popovski, and A. Dekorsy, "Massive machine-type communications in 5G: physical and MAC-layer solutions," *IEEE Communications Magazine*, vol. 54, no. 9, pp. 59–65, September 2016.

[123] Y. Yuan, Z. Yuan, G. Yu, C. h. Hwang, P. k. Liao, A. Li, and K. Takeda, "Non-orthogonal transmission technology in LTE evolution," *IEEE Communications Magazine*, vol. 54, no. 7, pp. 68–74, July 2016.

[124] C. Wei, H. Liu, Z. Zhang, J. Dang, and L. Wu, "Approximate message passing based joint user activity and data detection for NOMA," *IEEE Communications Letters*, vol. PP, no. 99, pp. 1–1, 2016.

[125] B. Wang, L. Dai, Y. Zhang, T. Mir, and J. Li, "Dynamic compressive sensing-based multi-user detection for uplink grant-free NOMA," *IEEE Communications Letters*, vol. 20, no. 11, pp. 2320–2323, Nov 2016.

[126] M. Z. Win, P. C. Pinto, and L. A. Shepp, "A mathematical theory of network interference and its applications," *Proceedings of the IEEE*, vol. 97, no. 2, pp. 205–230, Feb 2009.

[127] J. G. Andrews, F. Baccelli, and R. K. Ganti, "A tractable approach to coverage and rate in cellular networks," *IEEE Transactions on Communications*, vol. 59, no. 11, pp. 3122–3134, 2011.

[128] A. Rabbachin, T. Q. Quek, H. Shin, and M. Z. Win, "Cognitive network interference," *IEEE Journal on Selected Areas in Communications*, vol. 29, no. 2, pp. 480–493, 2011.

[129] A. Ghosh, N. Mangalvedhe, R. Ratasuk, B. Mondal, M. Cudak, E. Visotsky, T. A. Thomas, J. G. Andrews, P. Xia, H. S. Jo *et al.*, "Heterogeneous cellular networks: From theory to practice," *IEEE Communications Magazine*, vol. 50, no. 6, 2012.

[130] T. D. Novlan, H. S. Dhillon, and J. G. Andrews, "Analytical modeling of uplink cellular networks," *IEEE Transactions on Wireless Communications*, vol. 12, no. 6, pp. 2669–2679, 2013.

[131] T. Manabe and H. Takai, "Superresolution of multipath delay profiles measured by PN correlation method," *IEEE Transactions on Antennas and Propagation*, vol. 40, no. 5, pp. 500–509, 1992.

[132] C.-P. Li and W.-C. Huang, "A constructive representation for the fourier dual of the zadoff–chu sequences," *IEEE Transactions on Information Theory*, vol. 53, no. 11, pp. 4221–4224, 2007.

[133] J.-C. Belfiore, G. Rekaya, and E. Viterbo, "The golden code: a 2/spl times/2 full-rate space-time code with nonvanishing determinants," *IEEE Transactions on Information Theory*, vol. 51, no. 4, pp. 1432–1436, 2005.

[134] G. T. Buračas and G. M. Boynton, "Efficient design of event-related fMRI experiments using m-sequences," *Neuroimage*, vol. 16, no. 3, pp. 801–813, 2002.

[135] M. Wildemeersch, T. Q. Quek, C. H. Slump, and A. Rabbachin, "Cognitive small cell networks: Energy efficiency and trade-offs," *IEEE Transactions on Communications*, vol. 61, no. 9, pp. 4016–4029, 2013.

[136] D. Tuninetti and G. Caire, "The throughput of some wireless multiaccess systems," *IEEE Transactions on Information Theory*, vol. 48, no. 10, pp. 2773–2785, 2002.

[137] H. A. David, *Ordered Statistics*, 2nd ed. New York: Wiley, 1981.

[138] M. Shirvanimoghaddam, M. Condoluci, M. Dohler, and S. Johnson, "On the fundamental limits of random non-orthogonal multiple access in cellular massive IoT," *arXiv preprint arXiv:1705.10471*, 2017.

[139] A. Adhikary, X. Lin, and Y.-P. E. Wang, "Performance evaluation of NB-IoT coverage," in *IEEE 84th Vehicular Technology Conference (VTC-Fall)*, 2016, pp. 1–5.

[140] Y.-P. E. Wang, X. Lin, A. Adhikary, A. Grovlen, Y. Sui, Y. Blankenship, J. Bergman, and H. S. Razaghi, "A primer on 3GPP narrowband internet of things," *IEEE Communications Magazine*, vol. 55, no. 3, pp. 117–123, 2017.

[141] S.-M. Oh and J. Shin, "An efficient small data transmission scheme in the 3GPP NB-IoT system," *IEEE Communications Letters*, vol. 21, no. 3, pp. 660–663, 2017.

[142] T. Yang, L. Yang, Y. Guo, and J. Yuan, "A non-orthogonal multiple-access scheme using reliable physical-layer network coding and cascade-computation decoding," *IEEE Transactions on Wireless Communications*, vol. PP, no. 99, pp. 1–1, 2017.

[143] S. Verdu, "The capacity region of the symbol-asynchronous gaussian multiple-access channel," *IEEE Transactions on Information Theory*, vol. 35, no. 4, pp. 733–751, Jul 1989.

[144] Y. Lin and W. Yu, "Downlink spectral efficiency of distributed antenna systems under a stochastic model," *IEEE Transactions on Wireless Communications*, vol. 13, no. 12, pp. 6891–6902, 2014.